Vol. 44, No. 1 January, 1947

Psychological Bulletin

THE "LAWS" OF RELATIVE VARIABILITY OF MENTAL TRAITS

ROBERT S. ELLIS
Pomona College

The purpose of this paper is to survey the present status of our knowledge of the relative variability of mental traits and especially to examine critically the various generalizations and "laws" that have been proposed by writers on this subject.

The usually accepted method of computing relative variability is to use the formula devised by Karl Pearson for the Coefficient of Variation (V), according to which $V = 100\ S.D./M$. We divide the standard deviation (S.D.) of the distribution by the mean (M) of the distribution to two decimal places and by multiplying by 100 we eliminate the decimal. This makes it possible to compare the variabilities of characteristics which have been measured in different units such as pounds, inches, seconds and the unit scores of some psychological tests. It also makes an allowance for differences in averages when measurements are in the same units.

If we wish to apply the above formula to psychological test results we encounter in many cases a serious difficulty which is overlooked in some discussions of the subject. In order to use the V formula for most comparisons the measurements used must be arrived at by using scales with true zeroes and the scales must be in approximately equal units. It is especially important that V scores not be calculated for error scores or time scores which decrease as performance improves. In these cases zero scores would mean perfect performance rather than the absence of any ability.

Unfortunately the zeroes of psychological tests are commonly not true zeroes of ability. As a result the mean is usually too small, so that when we divide S.D. by M, the resulting V is too large. If then we compare V's on two tests which have zeroes at relatively different distances from their true zeroes the results cannot be interpreted accurately. This error is most serious when dealing with higher and more complex functions such as memory and reasoning.

While the Pearson method of computing relative variability has been generally used, other methods have had supporters. Thorndike (**86**, p. 9) suggested that the square root of M be used instead of M in the Pearson formula. This change increases relative variability in some cases where the use of M gives decreases. Yule (**109**, p. 49) suggests that deviations be compared by determining their ratios to the geometric mean. Wechsler (**98**) after eliminating pathological extremes determines the ratios of the highest to the lowest measurements. This method seems to have been used because of Wechsler's hypothesis that the ratio of the highest to the lowest measures was limited by the mathematical constant *e* (2.718..). This, however, does not seem to hold for some simple sensory functions. Woodrow (**105**) changes raw scores to scaled scores and compares S.D.'s to determine the effect of practice. Peters and Van Voorhis (**67**, pp. 78 ff.) throw the baby out with the bath water: they suggest that the zeroes of all distributions be placed about 3 sigmas below the mean where the scores begin to diverge. In normal distributions this would make all V's equal to about 33 and hence useless for comparison.

These differences in method are important because the conclusions reached may depend on the method used. However, since most students accept the Pearson formula we shall use it as a basis for further discussion.

Historical Background

The earliest suggestions of laws of variation were, quite naturally, made by biologists, but some of these have psychological implications and should be mentioned.

In 1809 Lamarck (**48**) opened the question as to the natural causes of variations by advancing his well-known idea that organisms vary in response to environmental stimulation and in a way to meet environmental requirements. Characteristics become stronger through use or weaker through disuse, and these changes are transmitted to offspring.

Fifty years later Darwin (**12**) included a chapter on "Laws of Variation" in *The Origin of Species*. He accepts the Lamarckian idea about use and disuse but places less emphasis on it and emphasizes variation from other " . . . causes of which we are quite ignorant." Among his conclusions are these: rudimentary organs are more variable than ordinary organs; highly developed organs are unusually variable; specific characters are more variable than generic characters; and lower organisms are more variable than higher organisms.

Havelock Ellis (**22**), in 1894, developed and emphasized the idea that men are more variable than women. The evidence for this claim

rested largely on physical characteristics. Three years later, Pearson (66) examined at some length the problem of sex differences in variability in physical traits. He concluded that evidence was lacking for greater male variability. Pearson also concluded from the study of cephalic indices that civilized races are relatively more variable than primitive races.

In 1902 Vernon (95) published a very important biological study of variation with incidental psychological references. This was based on a careful statistical treatment of measurements. After studying variations in physical growth he formulated a tentative law as follows: "The variability of a developing organism diminishes regularly with its growth" (p. 206). Minot and Pearson, working independently, apparently did the most important work which furnished the basis of this law. Vernon agrees with Minot that this reduction in relative variability during growth probably occurs in all mammals. Psychologists naturally will ask: Is this true for mental traits?

Vernon also discussed the possible influences of genetic selection and of natural selection on variability. In both cases, relative variability would be decreased.

In 1895 Binet and Henri (3, p. 417) after reviewing studies made up to that time in the psychological field conclude:

> Among the results which emerge from all these studies we shall cite a few: The first, the most important of all, we believe, is that the higher and more complex a process is the more it varies from individual to individual: sensations vary from one person to another, but less than memory, memory of sensation varies less than memory of ideas, etc. It follows from this then that if one wishes to study the differences between two individuals it is necessary to begin with the highest and most complex processes and it is only secondarily that we need to consider the simple and elementary processes: this is, however, the opposite of what has been done by the great majority of authors who have treated this question. (Translated by the writer.)

This statement was made of course before the statistical treatment of mental measurements had become common. It is probably based primarily on observations of qualitative differences.

I have not seen the first edition of Stern's *Differentielle Psychologie* which appeared in 1900 but in the third edition (83, pp. 257 ff.) he accepts the conclusion of Binet and Henri that complex traits are more variable than simple ones. To this he adds the idea that there is a positive correlation between the time of appearance (phylogenetically and ontogenetically) of a trait and its variability. Hence traits appearing in adolescence would be more complex and more variable than those appearing in infancy. He expresses the idea that relative varia-

bility should be determined by finding the percentage which a deviation is of the mean.

A study made by F. L. Wells seems to have suggested to Stern the idea of comparing the variability of traits within an individual with the variability of a single trait in a group of individuals. He concludes that higher processes will vary more than lower processes in a single individual. Presumably this would mean, for example, that mathematical ability and linguistic ability would show more independent variability in the same person than would auditory and visual acuity.

Thorndike (87, p. 317) formulates two tentative conclusions about relative variability as follows:

1. The variations are, in general, greater in acquired than in original traits.
2. They are, in general, greater in traits peculiar to man than in traits characteristic of all mammals.

Studies of sex differences are reviewed for evidence of a difference in relative variability and Thorndike concludes that women are probably slightly less variable than men (pp. 193 ff.). Several studies made in the period from 1908 to 1914 on the effects of practice on individual differences are reviewed (pp. 304 ff.) but the results are inconclusive. However, he at least formulates the problem.

H. L. Hollingworth (41, pp. 74–84) discusses the problem of relative variability and concludes that variability is probably greater for traits that are more complex, more functional, more recent (phylogenetically or ontogenetically), more specific, more symbolic, less used and less relevant.

Hull (42) makes a study of the question of variations in traits within the individual. He calls this "trait variability" and reaches a tentative conclusion that such trait variability is about 80 per cent of the variability of single traits in the population.

Wechsler (98) raises the question as to the range of human capacities. After eliminating the pathological extremes, he determines the ratio between the measurements of the highest and the lowest members of a group and draws the conclusion that with very few exceptions this ratio does not exceed 3.0:1 (p. 62). Wechsler emphasizes the point that while individual differences are real and are important they are not nearly as great as is commonly assumed. He believes this conclusion has an important bearing on the question of democratic government and on related social and economic problems.

Wechsler also subscribes to the idea that complex traits are more variable " . . . since the variability of any given phenomenon is neces-

sarily the function of the product of the variabilities of the individual factors which determine it" (p. 59).

This brief review will suffice to indicate the chief points in the historical background of current ideas of relative variability. Before attempting to pass judgment on the accuracy of these and other generalizations it will be useful to have a look at the problem of variation.

The Mechanism of Variation

Most normal psychological characteristics show continuous variation from the very low to the very high degrees of the trait. Geneticists

TABLE I

The Relative Variabilities (*V*) Of Distributions Resulting From Chance Combinations Of Coins And Dice

	S.D.	*M*	*V*	*S.D.*	*M*	*V*
4 Coins	1	6	17			
9 Coins				1.5	13.5	11
4 Dice	3.416	14	24			
9 Dice						
Theoretical				5.123	31.5	16.3
Empirical				5.073	32.3	15.7

Heads and tails on the coins are counted as 1's and 2's respectively Four coins falling three heads and one tail give a total of five.

The S.D.^{2} of the first N natural numbers, 6 in case of dice and 2 in case of coins as used above, equals $(N^2-1)/12$ (Yule, 109, p. 143). When we add equipotent independent variables the S.D.^{2} varies as the number (n) of the variables. Hence the S.D. of the distribution resulting from tossing n dice or coins is given by the formula: S.D.^{2} $=n(N^2-1)/12$. For nine dice this would be: S.D.$^2=9(6^2-1)/12=26.25$. S.D. $=5.123$. As an empirical check, nine dice were tossed 200 times with results as given under "Empirical."

usually interpret this to mean that differences in the strength of inherited characteristics are due to multiple factor inheritance: the strength of the characteristic is determined by the joint influence of a number of genes. This hereditary potential would be affected by environmental factors to a greater or lesser degree, so that the final result would depend on both kinds of factors.

In order to get examples of variation under conditions where we understand clearly the material with which we are dealing, some results, both theoretical and empirical, from certain chance combinations of coins and dice are presented in Table I.

When two factors combine to produce a more complex total the combination may be an additive process or it may involve multiplication with the result a product of the elements. Psychologists usually accept the additive assumption. The addition of normal distributions, as in case of our dice problem, gives a normal distribution while products give a skewed distribution. Since distributions of complex psychological traits are commonly approximately normal this seems to justify the tentative use of the additive method. The difference between these two assumptions, as we shall see later, is vitally related to the problem as to whether simple or complex traits are relatively more variable.

The "Laws" of Relative Variability

Complexity vs. Simplicity

If we examine Table I we find that, whether we are dealing with coins or dice, as we pass from the totals for four variables to nine variables the S.D.'s increase and the V's decrease. This happens because the M's increase in proportion to the number of variables while the S.D.'s increase in proportion to the square root of *N*. On this basis then an increase in complexity results in an increase in absolute variability but a decrease in relative variability.

If, however, we compare the totals for four coins and four dice we have the same number of variables in each case, but with an increase in the variability of the variables, and under these conditions, both the S.D.'s and the V's increase.

On this basis if we assume intelligence to be made up of capacities for perception, memory, reasoning and such other behavior as the reader wishes to add, the value of V for intelligence should be lower than the V's of its components. In another population where intelligence has the same components but where the components themselves varied more widely in strength the relative variability of intelligence would be greater than in the first population.

According to accepted statistical principles we may administer two tests to the same group of subjects and after computing the M's and S.D.'s of the two tests separately we may determine the mean of the sum of the two tests simply by adding the means of the two separate tests, and we may obtain the S.D. of the summed distribution by using the formula for that purpose (**109**, p. 211). From this it follows that the value of V for the sum of any two tests can never exceed the higher of the two values of V for the two tests which were combined—provided the test scores were of such a character that it was legitimate to compute a V in the first place. In this sense it is a statistical impossibility to in-

crease the value of V by combining test scores. If then complex traits are considered to be the sum of simple traits, they must be relatively less variable.

TABLE II

COEFFICIENTS OF VARIATION (*V*) FOR DIFFERENT CHARACTERISTICS

Trait	*Subjects*	
Weight of spleen (6)	54 normal males, 30–40 years old	45[1]
Olfactory acuity (4, 34)	Assorted adults	30+[2]
Visual acuity (20, 47, 77)	805 17–20 yr. olds plus others	27+
Strength of back (36)	Men 17–30 years old	22[3]
Simple visual reaction time (29, 35)	5564 English adults plus others	21[1,4]
Strength of grip (106)	609 16 year old boys	21[1]
Weight of liver (33)	73 English adults	20[5]
Auditory acuity (68)	61 young adults	19+
Vital capacity (106)	599 16 year old boys	19[1]
Card sorting (72)	148 16 year old girls	17
Visual perception span and speed (58)	98 males, 18–29 years old	15[5]
Memory span, digits (98)	236 male adults	15[5]
Tapping (106)	615 boys, 16 years old	13[1]
Highest audible pitch (47)	805 17–20 year olds	12
Body weight (43)	U. S. soldiers in 1917	12
Weight of cerebrum (65)	308 English adult males	11
Brain weight (65)	416 English adult males	9
Cranial capacity (51, 65)	English adult males	8
Knee height (43)	U. S. soldiers in 1917	8
Leg length (43)	U. S. soldiers in 1917	7
Stature (43)	U. S. soldiers in 1917	4
CAVD intelligence (88)	Assorted	3[5]
Body temperature (102)	601 convicts	0.5[5]

[1] M, S.D. and V calculated by the writer.
[2] V estimated by the writer from percentage distribution.
[3] V calculated by the writer from percentile distribution.
[4] Based on reciprocals of time scores.
[5] V calculated by the writer.

If we turn from statistical theory to empirical data to test our conclusion we confront the fact that most intelligence tests do not have true zeroes. Hence we cannot compute a true value of V for them. An exception is found in the CAVD test worked up by Thorndike and others (88). The mean adult score on this test is given as 36.5 and the S.D. is 1. This gives a V of 2.74. By examining Table II it will be seen that this is comparatively a very low value and agrees fully with theoretical expectations.

Even if we make a very large allowance for error in Thorndike's determination of the zero of the CAVD scale and multiply the value of V by 3 it would be only 8.2, which is relatively a low value.

Thurstone (**89**) also has attempted to determine the true zero for mental tests. His method is based on the assumption that relative variability of intelligence remains constant during growth. On a priori grounds the writer questions this assumption because all of the results he has seen for physical, psychomotor and sensory traits measured by objective scales with true zeroes show variations in V's with age. Under these conditions it would be very surprising if the V's for intelligence did not change with age. In the second place Thurstone's results are not consistent. One application of his method to Binet Test results (**89**, p. 196) gives a V of 7.2 (M = 13.9, S.D. = 1) while another application (**90**, p. 574) gives a V of 37.6 (M = 2.66, S.D. = 1). This difference can be only partly accounted for by the difference in the variability of the two groups tested. The growth curve in the first instance is a conventional negatively accelerated growth curve similar to those usually found. In the second instance the curve is first positively accelerated and later is negatively accelerated. These disagreements would seem to afford ample basis for questioning the validity of Thurstone's method.

Ability as represented by a total score is certainly more complex than the abilities represented by the separate test scores. Yet the V's for total scores, as noted above, are regularly lower than the V's of at least some of the individual tests. An example of this may be had from the VACO test results reported by Freeman and Flory (**25**, p. 42) for children aged 13 years. They report V's as follows: vocabulary 20.3, analogies 23.9, completion 20.3, opposites 27.9, and total 18.5.

If we consult Table II we find that with the exception of the weight of the spleen the highest V's are for olfactory and visual acuity, strength of back and of grip, and reaction time. These are hardly what psychologists would usually call the most complex traits. Visual perception span and memory span for digits are near the middle of the list.

Additional evidence that simple functions are relatively more variable than complex functions is provided by some studies not included in Table II.

Leukhart (**49**) gives the times for monocular accommodation for 14 subjects. Since higher time scores mean slower and poorer accommodation I have calculated the value of V for the reciprocals of these time scores and find it to be 40.6. Warden, Brown and Ross (**96**) determined the threshold for motion acuity for 28 subjects at scotopic

levels of illumination. This is stated in terms of angles and the lowest scores indicate the greatest visual acuity. For this reason I have again determined the V based on reciprocals and find it to be 147. Since one extreme case is responsible for a large part of this value, I determined V for the remaining 27 cases and found it to be 74—still a very high value. In both of the above studies the number of cases is rather small and the sampling is limited to groups of generally superior young adults. However, when we find wide variability under such conditions it appears improbable that we should not find it with a larger and less selected sample of the population.

Hermans (**38**) had 100 subjects observe the size of a standard 100 mm. aperture with binocular, with monocular and with pinhole vision and attempt to match this by adjusting a different aperture with binocular vision. Table III gives the results. M's and S.D.'s are from Hermans, V's have been added by the writer. Relative variability in responses is clearly greater with the simpler pinhole vision and less with the higher and more complex binocular vision.

TABLE III

RELATIVE VARIABILITIES OF JUDGMENTS OF SIZE AT THREE LEVELS OF VISION (HERMANS, **38**)

	M	*S.D.*	*V*
Pinhole vision	67.25	23.05	34
Monocular vision	93.70	14.30	15
Binocular vision	104.74	6.54	6

McGeoch (**52**) used three groups of nonsense syllables of different levels of association value, as determined by Glaze, and added a group of three-letter words. The 00% syllables are those for which subjects reported no associations, the 53% syllables are those for which 53 per cent of the subjects reported associations, and so on. McGeoch determined the amounts of these that were learned in a given period of time. Table IV gives the results. M's and S.D.'s are from McGeoch.

TABLE IV

THE RELATION OF RELATIVE VARIABILITIES TO THE ASSOCIATIVE VALUE OF SYLLABLES (MCGEOCH, **52**)

Material	*M*	*S.D.*	*V*
3-letter words	9.11	1.12	12
100% syllables	7.35	1.96	27
53% syllables	6.41	2.37	37
00% syllables	5.09	2.60	51

V's are added by the writer. This table shows that as material becomes more meaningful, and hence presumably involves more complex learning functions, there is relatively less variability in the amount learned.

While the zeroes of these scores may not be true zeroes, that can hardly account for the obtained differences. Unfortunately it is not possible to extend this type of analysis to a study of higher levels of rational memory because of the fact that nothing approaching a suitable measuring instrument is available.

In the psychological field the evidence certainly does not support the generalization that complex traits are relatively more variable—rather the opposite. This conclusion is further supported by considering anatomical measurements.

Some of the internal organs such as the spleen and the liver are relatively the most variable. If one questions these results because of admitted difficulties in securing a satisfactory selection of human bodies to dissect, it may be noted that the same relation is found for laboratory animals. Thus Donaldson (16, p. 225) reports V's for weights of the rat as follows: brain 10, body 19, heart 24, liver 25, thymus 34, and ovaries 43. As in man, body weights show less variability than is shown by some of its parts. The brain, which is hardly a "simple" structure, is relatively much less variable than the internal organs such as the spleen and the liver. Likewise, in Table II, consider the series—stature, leg length, and knee height. Stature is more complex and relatively less variable. Similarly the cerebrum is relatively more variable than the whole brain and the latter is relatively more variable than the brain case.

In contrast we may note that body weight is relatively more variable than stature. As a matter of geometry the body is a solid and has three dimensions. Similar solids vary as the cubes of their single dimensions—and cubes are *products* of three dimensions. In our dice and coin illustration we worked with *sums* and the results seemed to agree with the psychological test results. If, however, we deal with products the result is different and then Wechsler, as quoted above, would be right when he says that complex characteristics are relatively more variable "since the variability of any given phenomenon is necessarily the function of the product of the variabilities of the different factors which determine it" (98, p. 59).

While this generalization is true when we compare stature and weight it does not appear to be true when we compare psychological characteristics. Psychological test results agree better with the additive assumption usually made, and, as I shall point out later, the assumption of subtraction is helpful in understanding certain changes in

relative variability associated with old age, fatigue and forgetting. Certainly there is no good psychological, biological or mathematical reason why we must assume that psychological variation is always and "necessarily" a function of products rather than of sums or differences.

Functions vs. Structures

Hollingworth's second law states that functions are more variable than structures. From Table II we may see that the lowest V's are for functions while the highest is for a structure. With advancing age relative variability of visual acuity climbs rapidly and is above 90 by the age of 60 (**77**, p. 85). It might be difficult to match this with an example of structural variability. But in young adults and in children several structures are relatively more variable than visual acuity. Brain weight is relatively more variable than CAVD intelligence. A very wide range of V's can be supplied for both structures and functions, hence it would appear that this "law" hardly agrees with the facts.

Symbolic vs. Concrete

Since CAVD intelligence is distinctly symbolic and since the ability to handle symbols is presumably complex it hardly seems necessary to discuss this "law" separately.

Recent vs. Ancient

Stern, Thorndike and Hollingworth have supported the idea that traits recently acquired—whether phylogenetically or ontogenetically—are more variable than those of more ancient vintage. This seems to be in conflict with the principle supported by Vernon that the variability of a developing organism decreases with growth. Also it appears to be in conflict with the fundamental biological theory of evolution. Thus Shull (**81**, p. 243) writes as follows (italics are my own):

> The similarity of the species of a genus is held to indicate kinship, but since *there is greater diversity among the individuals of a genus than among the members of a species*, the common stock from which the species of a genus have sprung must have existed at an earlier time, in order that evolution could bring about the degree of divergence now observed.

Reptiles and lower animals are "cold-blooded" because they lack our temperature regulating mechanism. Yet our body temperatures show much less relative variability than various "older" functions. Also body temperature is less variable in adults than in children. The higher levels of CAVD intelligence are late in appearance both in the race and in the individual and show relatively small variation. The

sense of smell was originally the dominant distance receptor and the cerebrum first began to develop as an olfactory correlation center. Yet at the present time and in young human adults olfactory functions appear to be relatively more rather than less variable than visual and auditory functions. Auditory acuity develops phylogenetically after visual acuity and seems to be relatively less variable. Musical talent can hardly be said to appear early in the race but it does appear early in childhood and it seems to be more variable than abstract intelligence.

This generalization that traits more recently acquired—whether individually or racially—are more variable than older traits is probably a deduction from the supposedly greater variability of complex functions. It appears to be equally erroneous.

Specific vs. Generic

Darwin says: "It is notorious that specific characters are more variable than generic." And " . . . the points in which all the species of a genus resemble each other, and in which they differ from allied genera, are called generic characters." But since the nearest relatives of *Homo sapiens* are known only through fossil remains and since biologists are not agreed (Dobzhansky, **14**) on the classification of these fossils into genera and species it seems unwise to attempt to discuss this principle as Darwin has defined it.

Hollingworth's statement of the principle is that specific and less widespread traits are more variable than those that are more generic. Interpreted in this way it means about the same thing as Thorndike's second law that variability is greater in traits peculiar to man and as Pearson's claim that civilized man is relatively more variable than primitive man. It is also related to the question discussed above under "Recent vs. Ancient."

Among the important differences between man and our nearest surviving relatives, the gorilla and the chimpanzee, are man's upright posture, longer legs than arms, more uniform teeth, virtual absence of hair from most of the body, smaller face, and greater brain and intelligence.

Young human adults do not appear to vary greatly in upright posture when standing or walking. Schultz (**78, 80**) finds less relative variability in face, head, trunk and limb measurements in man than in some lower primates. Some primates are relatively more variable in arm length than man is in leg length.Relative absence of hair from the human body and limbs is a striking case of uniformity in a specific character. This is especially true of the human back. We have already

noted that the human brain and CAVD intelligence are not strikingly variable. On the other hand, our internal organs are generic and some of them are very variable. Visual and olfactory acuity are quite generic and are also quite variable.

All members of the *Order primates* are alike in having one head, one trunk, four limbs, two ears, one tongue, etc., but if we are studying quantitative variations in characteristics—and that happens to be our present problem—we find some of the highest coefficients of variation for characteristics which are common not merely to the members of a genus but to the members of an entire family, order or class. The distance between the nipples of human females has a V of about 20 (92, p. 108). This is a mammalian characteristic and is much higher than the V's for most linear measurements. Wide variations do occur in a single species, especially in domesticated animals which have been subjected to controlled breeding and selection, but even here the members of a breed of dogs or chickens tend to be alike rather than different. However, if less widespread traits are more variable than more generic traits it would seem that a single breed of dogs or chickens should show very wide variation in those characteristics which are peculiar to the breed.

Less Relevant vs. More Relevant

Some writers have held that the variability of traits is inversely correlated with their biological relevance. This means that the less important the trait is for survival the more it may be expected to vary in the species. This seems to be based on the apparently reasonable deduction that the greatest variability cannot exist in things which are closely related to life. Against this is the fact that some of the internal organs are highly variable. Sensory capacities are essential for environmental adjustments but some of them are very variable. In lower animals many examples could be given of characteristics which seem to have little or no survival importance but which are rather uniform in a species. "A survey of the characters which differentiate species (and to a less extent genera) reveals that in the vast majority of cases the specific characters have no known adaptive significance" (Robson and Richards, 76, p. 314). Yet it seems obvious that if these characters were not reasonably uniform in a species they could not be used to differentiate species.

Less Used vs. Most Used

According to Hollingworth (41, p. 78), "Infrequently used traits are more variable than are traits or activities more constantly em-

ployed." As a psychological principle we might make something of a case for this "law" in connection with the acquisition of skills in the individual. This problem will be discussed under the effects of practice.

The Effects of Practice

Thorndike formulated and defended the idea that acquired traits are more variable than native ones. This is tied up with a lively controversy over the question as to the effects of the environment, including education and practice, on individual differences. Given certain individual differences, how will these be affected by adding certain variables in the form of practice or training?

If we consider the total "scores" arrived at by tossing n dice to represent ability before training and that the effects of practice could be represented by adding to these original scores some new scores secured by tossing p additional dice, each child's achievement after practice, could be represented by the total of $n+p$ dice. If this situation holds it follows from previous discussions that the normal effect of education and practice is to reduce relative variability.

We have evidence, however, that our n and p variables in school work are not independent variables but are negatively correlated. The dullest students in school are prodded most and work hardest. Frequently the brightest students are permitted to loaf. Thus May (55) in a study of college students found a correlation of $-.35$ between intelligence and the time spent in study. Drake (17) gave tests to college classes in biology and in history at the beginning and at the end of the semester, converted the raw scores to standard scores, and subtracted scores on the first test from scores on the second test to determine gains. He found negative correlations between first scores and gains and between intelligence and gains. Kelley (45) challenged the then current conception and held that formal education especially at the elementary school level reduced individual differences in several respects. Meltzer and Bailor (56) tested psychology students at the beginning and end of a course and found men decreased from 40 to 21 and women from 49 to 14 in relative variability in knowledge of the subject

One method of studying the effects of practice and training is to compare some of the results secured from intelligence tests and achievement tests. If these results are expressed in terms of age or grade norms and the range from the tenth to the ninetieth percentile points is determined we find that variability in achievement is usually less than variability in intelligence. For example, at the age of ten years the

tenth to ninetieth percentile range on the revised Stanford-Binet test is 4.15 years (**85**, p. 40). On the Stanford Achievement Test, Cornell (**11**, p. 87) finds the corresponding range to be 3.4 years. This method largely avoids the difficulty imposed by the fact that the tests do not have true zeroes. The results do not justify the conclusion that ordinary school training increases individual differences within the trained group.

In controlled practice experiments gross gains are usually positively correlated with initial scores, but percentage improvement and initial scores are negatively correlated. Individuals with lower scores make relatively larger gains than those with larger scores.

Kincaid (**46**) reviewed 24 experiments on the effects of practice and found negative correlations between initial scores and percentage improvement in 22 of the studies. In 19 of the 24 experiments larger S.D.'s were found after practice than before, but in 16 of the 24 cases V's were smaller than before. Both the S.D. and the M usually increase with practice but the latter increases more, with the result that relative variability usually decreases.

Reed (**73**, **74**), reviewing practice studies to 1931, concludes that in 77 per cent of the studies ($N=70$) V decreases with practice, and in 93 per cent of the studies ($N=58$) the correlations between initial performance and per cent improvement are negative.

Burns (**8**) reviews 84 practice studies and agrees essentially with Reed in his findings. He suggests that one of the important reasons for differences in the results of practice studies is to be found in differences in motivation.

Additional practice studies have been made since 1937 but they have not changed the general picture nor have they made it clear why in some cases relative variability increases with practice. A recent study by Yoshioka and Jones (**108**) of stylus maze learning reports that with practice there is a sharp increase in relative variability in errors but not in time scores. In the latter case the work unit was constant and time scores decreased as learning progressed. In both cases then V's have been computed for scores for which zero would be a perfect score (though impossible in case of time scores). Yet, as was noted earlier, we cannot legitimately use V scores unless the scale has approximately at least a true zero.

A hypothetical example will clarify the difficulty. Children given a spelling test have an average of 80 right, 20 wrong, S.D. = 5. After practice they test 96 right, 4 wrong, S.D. = 2. V's based on "rights" are 6 before practice and 2 after practice. Based on "errors," V's are 25 before practice and 50 after prac-

tice. Clearly the true variability in knowledge of spelling is not represented by a V of 50 after practice. For this reason before V's are computed all scores which show improvement by decreasing scores must be converted to a form in which increasing ability is represented by increasing scores. Otherwise the V scores may show the opposite of the true situation, though V's based on time scores are usually more nearly correct than V's based on error scores.

Relatively few experiments have shown statistically significant increases in V's after practice. Further study of the exceptions is needed to explain why they give different results. If we took 100 individuals, divided them into four equal groups of equal initial ability and then gave distinctly different amounts of practice to each group, relative variability for the total group would normally be greater after practice then before. When environment operates differentially in this way, as it apparently does at times, relative variability might be increased. Those with higher intelligence tend to remain in school longer so that amounts of practice are distinctly unequal. This should produce a wide spread in achievement along particular lines. But to measure this fairly we must have tests with true zeroes and with equal units.

Forgetting

To a considerable degree forgetting is a reversal of the process of learning. In terms of our dice illustration it is possibly similar to what would happen if we started with "scores" based on throwing n dice and subtracted from each score the scores resulting from throwing m dice, m being less than n. Apparently there is a low negative correlation between learning capacity and rate of forgetting. The mean score after forgetting would be lower than the original score and the standard deviation would be larger (**109**, p. 211). From this it follows that if the dice illustration holds, we should expect an increase in relative variability after forgetting.

Tilton (**91**) reviewed 39 studies of the effects of forgetting on individual differences and found that in 24 cases absolute S.D.'s increased. The ratio of the average S.D. after forgetting to the average S.D. before forgetting was 106.7 to 100. The average V after forgetting divided by the average V before forgetting gave 165.6. In the majority of the cases then experimental evidence has shown that forgetting does increase relative variability. Forgetting tends to return a group to the greater relative variability which existed before practice.

Watson (**97**) reports that after forgetting relative variability is greater for recall than for recognition. Probably this is related to the fact that power of recall is lost more rapidly than power of recognition.

Effect of Fatigue

Fatigue, like forgetting, involves a loss of function. It differs in that it is more immediately and acutely related to physiological functions such as circulation and respiration. Short and long work periods may differ because the short work period may not be greatly influenced by "second wind" while the long experiment is so influenced.

Wells (**100**) had 10 subjects tap for 30 seconds and compared the first 5 seconds with the last 5. Under these conditions, with a short work period, a reduction in relative variability was found.

With long work periods, increases in relative variability have usually been found. Weinland (**99**) had 10 subjects work with the ergograph over a period of six months and reported that V increased with fatigue. He concluded that increased variability was due largely to loss of control. Manzer (**54**) had 27 college men work to exhaustion on the ergograph and found that relative variability of work with fatigued muscles was 309 per cent of what it was with unfatigued muscles. Philip (**69**) had twelve subjects work about seven hours at tapping until they were very fatigued. He found an increase in relative variability with fatigue and like Weinland attributed this largely to loss of control. Edwards (**18**) reports the effects of the loss of 100 hours of sleep on 19 subjects. He finds no great differences in V's for controls and experimental group but does find that the extremes in reaction times are in the experimental group. Flugel (**24**) had 46 children do addition for 20 minutes daily for 46 days. He reported that fatigue and ability were found to be negatively correlated. This should give an increase in relative variability.

Some studies have reported qualitative changes such as doing the right thing at the wrong time. Marked increases in irritability are also reported. The unfatigued individual is evidently more stable and predictable.

One of the technical difficulties in the study of fatigue is found in the fact that long fatigue tests are monotonous. Hence what is classified as fatigue is probably partly boredom and lack of effort. This factor would tend to produce variation in the results of different experiments.

Age Differences

Vernon's theory that variability decreases with growth has been mentioned. V's based on data from Montague and Hollingworth (**59**) for length and weight of infants at birth are about 6 and 15 respectively. For young adults the corresponding figures are about 4 and 12, based on Army results for men (Table II) and on data from Doll (**15**, after

Smedley) for women. Thus far then the principle seems to hold. However, there is a prepubertal acceleration in growth which results in a rise in V's at that period. The greatest prepubertal or pubertal relative variability in boys occurs at about 14 years and in girls at about 12 years. After birth then, V's for height and weight fall, then rise, then fall to maturity. In young adults relative variability is lower than at birth.

Sensory and psycho-motor functions follow varying patterns with respect to relative variability during growth. No simple generalization seems to cover them. Relative variability in visual acuity rises from about 12 years to maturity. Relative variability in throwing a ball declines during adolescence in boys but seems to increase slightly in girls.

Henmon and Livingstone (37) collected data from different sources and studied changes in relative variability of psycho-motor and mental functions during growth. They found relative variability to decline with age to maturity, and did not find any consistent evidence that relative variability increases at the prepubertal period.

Israeli (44) finds decreasing variability in aesthetic judgments with age to maturity. Miles (58, after Price) reports data on visual perception, span and speed which show a decline in V's with age to maturity.

V's based on group intelligence test results generally show declines with age to maturity. Freeman and Flory (25) report the results of a ten year study with the VACO Tests. This covers ages 8–17 inclusive. V's decline regularly with age without evidence of a prepubertal bulge. Lincoln (50) reports V's for the Yerkes-Bridges, the Pressey, and the Dearborn intelligence tests. All show declines with age. Odom (60) collected data for the Otis, the National and the Illinois intelligence tests. They also show declines in V's with age. Adkins (1) finds that on retests with the Otis and with the Morgan tests V's decline from grade 7 through grade 12.

The fact that these tests do not have a true zero leaves some doubt as to the exact significance of the results reported. However, analysis of mental growth curves has convinced the writer that there is a real decrease in the relative variability of intelligence during adolescence.

The Stanford-Binet is not a suitable instrument for measuring changes in relative variability with age. It is used on the assumption that the S.D.—and hence V—remains constant at 16. Yet the original report by Terman and Merrill (85, p. 40) and later studies by Goodenough (30) and Brown (7) show marked declines in S.D.'s from 2.5 to 6 years, then a marked rise to 12 years, and a decline thereafter. This is a prepubertal bulge in relative variability, but whether it is due to

variations in growth or to the nature of the test is not clear. This is of practical importance because it means that a child who is 2.5 S.D.'s above average may flucutate 20 points in IQ simply by following the normal growth pattern for such children.

The available evidence indicates that relative variability of intelligence declines during adolescence, but we cannot yet state with any certainty the exact changes that occur from birth to maturity.

After maturity the general tendency is for relative variability to increase. Tests with an important speed factor show marked declines in the averages with age while those which emphasize power rather than speed are on the average less affected. However, Gilbert (**28**) reports losses in memory in senescence which show V's up to 87. This is for retention of Turkish-English vocabulary. For the age period from 20 to 29 years the corresponding V is 23.

The results secured by Miles and Miles (**57**) which showed a decline in intelligence in old age were secured with a speed test. On this the S.D.'s change little from 30 to 70 years but the mean scores decline greatly and V's increase correspondingly. In another study already referred to above, Miles (**58**) submits data from Price showing M's and S.D.'s from age 6 to old age. For visual perception, span and speed at 6–7 years $V=32$, at 20–24 years $V=13$, and for 75–79 years $V=34$. Goldfarb (**29**) found an increase in relative variability in reaction times in males in old age but the changes in females were not reliable. The marked increase in relative variability in visual acuity with age has already been referred to. V rises above 90 at the age of 60 in men (**77**, p. 85). The V for auditory acuity also rises with age but to a lesser degree.

There appears to be a fairly definite tendency for relative variability to increase after early maturity, but there are great differences in the changes in different functions. An extreme increase is found in visual acuity while strength of grip shows only a slight increase.

Variability and Change

Darwin held that both rudimentary and highly developed structures are unusually variable. Schultz (**79**, p. 321) makes a related suggestion: "Upon finding such a high variability in the human ear we are justified in suspecting that this structure is undergoing some evolutionary change, in the form of either an increase or a decrease in size." These conclusions, added to the finding of differences in relative variability associated with age changes, seem to justify this tentative conclusion: *The relative variabilities of both structures and functions tend*

to be positively correlated with rates of change, both phylogenetic and ontogenetic, and both progressive and regressive.

Schultz cites human wisdom teeth and little toes as examples of widely variable structures which seem to be in process of regression. We found sight and smell relatively more variable than hearing. Smell has been undergoing involution while sight has been evolving.

This hypothesis agrees well with what is known about individual growth and decline. Quantitative individual differences are produced by differential growth rates—of which the IQ is one index. Relative variability is high in the foetus and in infancy when growth is most rapid. Growth rates and relative variabilities both decline with age until the prepubertal period. At this point physical growth is accelerated and the corresponding V's rise. From this point, both growth rates and V's decline to maturity. Measurements of mental growth have not shown the prepubertal acceleration found for physical growth, and V's for mental functions do not rise at this point. As mental functions begin to decline with advancing age the corresponding V's increase. Visual acuity declines very rapidly between 40 and 60 years and V rises above 90.

High relative variabilities connected with change can in part at least be attributed to differences in the timing of changes in rates of growth and decline. Richey (75, p. 67), speaking of physical growth during the prepubertal and adolescent periods, comments:

> In general, it may be stated that measures of variability increase during the period that growth is comparatively rapid. A large part of the variability found for any particular group is probably due to differences in the periods of acceleration and retardation of the growth rates of the individuals making up the group.

The prepubertal spurt in physical growth occurs earlier in some children than in others, even among those who will as adults be the same height. And decline in old age occurs earlier in some than in others. There are also differences in the age at which growth stops. Altogether these differences in rates of growth and decline and in the timing of changes of rate are responsible for a large part of variability.

Sex Differences

Scientific males have fairly commonly credited males with greater variability and have used this to explain the larger number of male geniuses.

Pearson (66) assembled a considerable mass of statistics dealing with sex differences in physical measurements and concluded that human fe-

males are more variable than males. Pearl (**65**) collected data for about 40 comparisons of the sexes on physical measurements and his results show that V's for females are higher in a ratio of about 3 to 2. Todd and Lindala (**92**) made about 60 measurements of both white and Negro males and females, a total of about 120 comparisons. V's for females are larger in a ratio of about 3 to 1. From these results it seems necessary to conclude that *adult* human females are relatively more variable than adult males in the majority of physical measurements. However, the average difference is usually small.

It is generally accepted that females mature physically earlier than males. The average difference in age at the time of arriving at maturity is probably two or three years. And in discussing age differences we concluded that, with the exception of the prepubertal period, relative variability in height and weight at least decline with age to maturity. It follows that if we compare the heights and weights of boys and girls of the same age the girls are likely in the majority of comparisons to be relatively less variable simply because they are more mature. Their lower relative variability is a mark of greater maturity and not of a sex difference in relative variability. This possibility is recognized by Lincoln (**50**, p. 164) who says that variability may be a function of maturity. Since most of our statistics have been of school children we seem thus to have succeeded in reversing the true picture and have erroneously concluded that males in general are relatively more variable than females. If it were not for sex differences in the ages at which prepubertal and pubertal accelerations in growth take place, it is probable that males would more consistently be relatively more variable than females during the growth period.

On intelligence tests at the adult level we do not know accurately how the sexes compare either as to the average level of ability or as to variability. Prevailing opinion is that the average level is about the same, but that there are average differences in ability to score on specialized tests.

That there is a sex difference in intelligence at ages 11, 12 and 13 is shown by Table V. Scores on different tests have been made comparable by this method: the test score for 11-year-old girls is changed to 11 and the test score for 13-year-old girls is changed to 13. The other scores are then converted to this scale by proportion.

Terman (**84**, p. 66) does not report the numbers of cases for each age and sex but he gives a total of 905 cases for ages 5 to 14 inclusive. This is an average of 45.25 cases for each age and sex group. Hence I have used 45 in the table. The numbers of cases given for Pyle's test

TABLE V

SEX DIFFERENCES IN INTELLIGENCE AT 11, 12 AND 13 YEARS AS DETERMINED BY NINE INTELLIGENCE TESTS

Test		*Age 11*		*Age 12*		*Age 13*	
		N	*Mean*	*N*	*Mean*	*N*	*Mean*
National (101)	B	613	10.08				
	G	643	11.00				
Illinois (101)	B	155	10.51	171	11.27	166	12.13
	G	175	11.00	184	12.19	157	13.00
Stanford-Binet, 1916 (84)	B	45?	10.27	45?	11.71	45?	12.83
	G	45?	11.00	45?	12.03	45?	13.00
National (53)	B	98	10.33	89	11.69	89	12.16
	G	88	11.00	98	11.96	67	13.00
Pyle (71)	B	73	9.71	80	10.42	82	11.44
	G	73	11.00	89	12.07	73	13.00
McCall (9)	B	94	10.80	102	11.18	97	12.10
	G	98	11.00	101	11.82	84	13.00
Pressey (70)	B	179	10.41	182	11.26	174	12.48
	G	167	11.00	180	11.76	174	13.00
Army Alpha (10)	B	22	10.44	34	11.55	35	11.53
	G	33	11.00	31	12.12	35	13.00
VACO (25)	B	150	10.83	163	12.01	149	12.83
	G	176	11.00	168	12.03	144	13.00
Yerkes *et al.* (107)	B	33	10.91	34	12.11	28	12.75
	G	29	11 00	32	12.68	26	13.00
Weighted Means	B	1453	10.38	900	11.42	865	12.29
	G	1527	11.00	928	12.01	805	13.00
Adjusted Weighted M's	B			3218	11.367		
	G			3260	12.000		
Difference and standard error					.633 ± .0464		

are averages based on his tables for seven different tests. The score for boys on the National Test reported by Whipple is arrived at by counting 15 points as 1 year—this from the test norms. On the Yerkes Point Scale, all cases from 11 to 12 are grouped at 11, and so on. The final difference score, called the adjusted weighted mean difference is found by adding 1 year to the 11 year scores and by subtracting 1 year from the 13 year scores. This groups all cases at 12 years. Distribution S.D.'s are assumed to be 16 IQ units in standard error calculations. Hence for 12 years the S.D. used is 1.92.

This method gives the boys a mean score of 11.367 and the girls a mean score of 12. The difference, .633, is equal to 7.6 months. Taking three standard errors below and above the difference gives us .494 and .772. These are equivalent roughly to six months and nine months. Hence we conclude that for the above battery of tests there is a sex difference in mental maturity of from six to nine months at the age of twelve years.

Something should be said about results not included in the table. Whipple reported National test results for four cities for children aged 11 years. In two of these cities a special effort was made to test every child of this age. This apparently was not done in the other two cities. I have included the two cities where all cases were tested and have excluded the others. Wechsler-Bellevue and Revised Stanford-Binet results are not included because an effort was made to eliminate sex differences when the tests were standardized. Dearborn Test results are excluded because Lincoln (**50**, pp. 166 ff.) has shown that the test content unduly favors boys. As a result, it, like the two tests just mentioned, gives about equal scores for the sexes.

A battery consisting of these three tests would show the sexes about equal So far as I know, no similar battery of general intelligence tests now in use in this country will show boys mentally more mature than girls at these ages. At the high school level, boys are likely to test higher than girls because the boys are more drastically selected.

If girls test higher than boys at age 12, and if the sexes arrive at the same final average level, this shows that the girls are maturing earlier. Since we found that relative variability decreases with age to maturity it seems that we confront about the same situation with respect to intelligence as that found for physical traits. When age comparisons show girls with lower V's, the correct interpretation seems to be that girls are more mature, and. being more mature, they are consequently relatively less variable. When the boys arrive at the same maturity level, the V's for them will have decreased. To ignore the maturity

TABLE VI

RELATIVE VARIABILITIES OF THE SEXES ON TESTS IN WHICH ONE SEX IS DEFINITELY SUPERIOR

Test	*N*	*Sex*	*M*	*S.D.*	*V*
Artificial Language (19)	3236	M	22.8	11.0	48
	2632	F	29.5	12.9	44
Arithmetic (19)	3236	M	29.6	12.5	42
	2632	F	21.9	11.2	51
English (19)	3453	M	49.9	18.1	36
	2880	F	55.3	18.4	33
Mathematics (19)	3453	M	34.7	14.9	43
	2880	F	25.2	12.3	49
Geometrical Constr. No. 8 (93)	540	M	10.7	3.4	32
	703	F	8.9	4.3	48
Tonal Intensity (53)	100	M	81.8	8.3	10
	100	F	77.3	11.3	15
Tonal Memory (53)	100	M	75.5	11.8	16
	100	F	59.7	16.1	27
Beta 2, Cube Analysis (103)	1161	M	13.12	4.80	37
	1160	F	10.06	4.08	41
Beta 5, No. Checking (103)	1161	M	24.42	8.15	33
	1160	F	27.72	8.70	31
Conservatism-Radicalism (82)	181	M	23.8	9.0	38
	206	F	26.9	8.4	31
Unpleasantness (Sense of) (82)	187	M	31.6	16.7	53
	202	F	34.6	15.8	46
Throwing a Ball (15 yrs.) (23)	117	M	134.8	24.18	18
	138	F	73.6	18.5	25
Grip Strength (22.5 yrs.) (77)		M	82.16	11.86	14
(20)		F	51.00	10.20	20
Mechanical Comprehension (2)	54	M	43.78	7.70	18
	53	F	27.57	8.83	32

variable and to draw conclusions about adult sex differences from age comparisons of boys and girls leads apparently to error, and possibly to the reverse of the truth.

We have found that relative variability usually decreases with development. On this basis, if females are better than males in Trait A and are poorer in Trait B, they should, other things being equal, be relatively less variable in A and more variable in B. The largest sex differences on the American Council Psychological Examination are found for artificial language and arithmetic. On the Iowa High School Content Test the differences are greatest for English and mathematics. Table VI supplies the California Junior College norms for these tests as reported by Eells (**19**). To these results I have added geometrical construction (Touton, **93**), tonal intensity and tonal memory (McNemar and Terman, after Church, **53**), Army Beta 2 and 5 (Winsor **103**), conservatism and sense of unpleasantness (Skaggs, **82**), throwing a ball (Espenschade, **23**), strength of grip (Ruger and Stoessiger, **77**; Elderton and Moul, **20**), and mechanical comprehension (Bennett and Cruikshank, **2**). V's have been calculated or checked by the writer and are given to the nearest whole number.

The majority of these V's are too high because of the lack of true zeroes. V's for throwing a ball are too high because, as Hill (**40**) points out, the distance travelled by a projectile varies as the square of the initial velocity. Hence theoretically we should take the square root of the original measurements before computing the value of V. However, these defects affect both sexes in the same way, so that our comparison of the sexes is valid even though the V's are too high.

An examination of the table shows that the hypothesis holds in all cases. However, no claim is made that this is always true. It should be said that this table is made up of characteristics which show the largest sex differences in averages. If a random selection of traits were studied, most of the differences would be small, and as a result of errors of sampling and of measurement a much less consistent result should be expected. In any case, the hypothesis seems to work often enough to justify the suggestion that adult women will more commonly be relatively less variable in those things in which they are better; while men will be relatively less variable in those things in which they are better.

Absolute variabilities are less consistent. In six out of fourteen cases the higher mean goes with the lower S.D.; in the remaining cases the higher mean goes with the higher S.D.

As related to the question of genius, absolute variabilities seem

more important than relative variabilities, because, as was stated in the preceding sentence, in eight cases out of fourteen, the higher mean is associated with the higher S.D. However, Table VI is in most cases not based on the performances of mature adults, and measurements for making a fair comparison of adults are not available.

Race Differences

The V for *opinions* on this subject is very large, but the problem must remain unsettled until we have more measurements.

Differences within the Individual

Trait Variability. We have mentioned Hull's estimate (**42**) that on the average the different traits of a single individual vary 80 per cent as much as a single trait varies in a group of individuals. Instead of trying to answer the question in that form we can give a more certain answer if the question is put more specifically.

The amount of trait variability in the individual depends on the intercorrelations between the traits in question. Ghiselli (**27**) has supplied a formula for computing the extent of trait variability from the average intercorrelation between traits. Most investigators have found that the intercorrelations between simple motor functions are near zero. This means that trait variability of motor functions is about 100 per cent of the variability of a single trait in a group. This is verified in a study made by Owens (**61**). As the intercorrelations between the traits studied increase, trait variability within the individual decreases. Presumably the more nearly we are able to measure pure and independent traits the lower the intercorrelations will be and the greater the trait variability.

Measurements of complex functions, achievement tests included, usually involve considerable overlapping with the result that individual variability is correspondingly reduced. Gray (**31**) reports that the average range of individual variation on six achievement tests is two S.D.'s. This means that individual variation on such a limited battery of tests is about 35 per cent of group variation.

From the foregoing it appears that the amount of trait variability found in any particular study will depend largely on the battery of tests used. Any general average for all psychological traits would under present conditions be somewhat arbitrary and tentative.

Owens (**62**) studied the effects of practice on a group of motor traits and found that they did not become more alike with practice. This seemed to indicate that motor trait differences were due to innate causes.

The question has been raised as to whether there are differences in trait variability at different levels of ability. Among the studies of this problem are those by De Voss (13), Hertzman (39), Bown (5), and Gray (31, 32). Their results are conflicting and inconclusive. However, Garrett (26) has shown that the correlations between such functions as verbal, numerical and spatial abilities decrease with age to maturity. From this it would follow that the amount of trait variability would be positively correlated with the level of ability. The failure of other workers to find clear evidence of this trend may be attributed to either or both of two factors: the abilities tested were too complex or the age range of subjects was too limited.

Quotidian Variability. Woodrow (104) uses the term "quotidian variability" to cover variations in level of performance from day to day. Owens (61) uses the term "repetitive variations" and includes under this the systematic changes due to learning. In a study of motor skills Owens finds repetitive variations to be about 13 per cent as great as individual differences, and he attributes 90 per cent of the repetitive variations to learning. This makes quotidian variability unimportant in his study.

Elliott (21) finds that strong motivation decreases variability of performance on practice tests in rats. From this we can infer that a part at least of quotidian variability is due to differences in concentration and effort. This, of course, is what we should expect.

Variations in the results of achievement, intelligence, personality and other tests show that individuals do vary considerably from day to day. The unreliability of questionnaire tests is well known. So is the unreliability of teachers' marks. All of these depend largely on quotidian variability.

Paulsen (63) points out that coefficients of reliability obtained by the split-half technique are higher than those obtained by the test-retest method. This difference is due to fluctuations in the strength or efficiency of the traits themselves. He proposes to measure this by correcting the test-retest reliability for attenuation. This will indicate the highest possible test-retest reliability The Spearman-Brown correction formula does not apply to test-retest reliabilities, and in a later study of steadiness (64), Paulsen finds that the highest possible test-retest reliability is about .80. This means that the factors contributing to quotidian variability in steadiness are much more important than Owens found in his study of motor abilities.

A related study of trait consistency on the behavior side has been made by Trawick (94). He finds consistency in performance to be an

important indicator of the integration of personality. The more consistent individuals are generally more self-confident, ascendant, have higher self-esteem, are more goal-seeking, more sociable, more predictable, objectively more modest, and have more social insight. From this point of view, the study of quotidian variability involves us in the problems of unstable personalities.

This should be a profitable field for future study and particularly do we need to learn more about the limits of test-retest reliabilities.

Discussion

If we arbitrarily define a superior person as one who scores three sigmas above average, the superior person is about twice as good as the average in relatively simple capacities such as visual and olfactory acuity. As we descend the V scale the superior person's ratio of superiority decreases. On tapping speed the superior person is about forty per cent better than average. On stature he is only about twelve per cent above average. On level of CAVD intelligence he is only about eight per cent above average. Also if total ability is considered to be the sum of different more specialized abilities, the superior individual will deviate less from the average in total ability than he does in some specialized abilities. On this basis probably we should not be surprised when we find that a successful politician shows more superiority in the particular characteristics required for winning votes than he shows later in the way of general administrative ability.

General Conclusions

Much of the confusion found in conflicting statements about relative variability is due to the use of different methods, to failure to discriminate between absolute and relative measures of variability and to the use of the V formula with scores that are far from having true zeroes.

Too many of the "laws" and broad generalizations about relative variability proposed by different writers will not stand critical examination. More complex, higher and more recently developed functions tend to be relatively *less* rather than relatively more variable. As judged by relative variabilities, complex mental functions seem to be the sums rather than the products of simpler functions.

Practice usually reduces relative variability. Fatigue and forgetting usually increase relative variability. Changes in relative variability with age differ greatly according to the trait under consideration but with a definite tendency for relative variability to decrease with age

until maturity and to increase thereafter. Sex differences in relative variability are small, but greater relative variability is more often found in boys. This probably means that girls of a given age are more mature than boys of the same age and not that females are generally relatively less variable than males. Adult females are relatively more variable physically than males. In general, low relative variabilities indicate superiority. Each sex tends to be relatively less variable in those traits in which it is superior.

Relative variabilities are determined largely by rates of growth and decline and by differences in the timing of changes in rate.

The per cent which intra-individual trait variability is of individual differences ranges from 100 downwards, depending on the traits tested. The study of quotidian variability is largely in the exploratory stage.

BIBLIOGRAPHY

1. Adkins, Dorothy C. The effects of practice on intelligence test scores. *J. educ. Psychol.*, 1937 **28**, 222–231.
2. Bennett, G. K., & Cruikshank, Ruth M. Sex differences in the understanding of mechanical problems. *J. appl. Psychol.* 1942, **26**, 121–127.
3. Binet, A., & Henri, V. La psychologie individualle. *Année psychol.*, 1895, **2**, 411–465.
4. Blakeslee, A. F. Tests of the sense of smell at the New York Flower Show. *Eugen. News*, 1935, **20**, 75-76.
5. Bown, M. D. Variability as a function of ability and its relation to personality and interests. *Arch. Psychol., N Y.*, 1941, No. 252.
6. Boyd, Edith. A method of establishing the probable limits of normal variation in the weights of organs. *Anat. Rec.*, 1935, **62**, 1–6.
7. Brown, F. The significance of the IQ variability in relation to age on the revised Stanford-Binet scale. *J. genet. Psychol.*, 1943, **63**, 177–181.
8. Burns, Z. H. Practice, variability and motivation. *J. educ. Psychol.*, 1937, **30**, 403–420.
9. Commins, W. D More about sex differences. *Sch. & Soc.*, 1928, **28**, 599–600.
10. Conrad, H. S., Jones, H. E., & Hsiao, H. H. Sex differences in mental growth and decline. *J. educ. Psychol.*, 1933, **24**, 161–169.
11. Cornell, Ethel L. The variability of children of different ages and its relation to school classification and grouping. *Univ. of the State of New York Bull., Educ. Res. Stud.*, 1937, No. 1.
12. Darwin, C. *The origin of species by means of natural selection.* (Amer. Ed.) New York: Appleton, 1865.
13. De Voss, J. C. Specialization of the abilities of gifted children. In L. M. Terman *et al., Genetic studies of genius*, Vol. 1, Ch. 12. Stanford: Stanford Univ. Press, 1925.
14. Dobzhansky, T. On species and races of living and fossil man. *Amer. J. phys. Anthrop.*, 1944, n.s. **2**, 251–265.
15. Doll, E. A. *Anthropometry as an aid to mental diagnosis.* Vineland,

N. J.: Vineland Training School, 1916.

16. DONALDSON, H. H. *The Rat.* Philadelphia: Wistar Institute, 1924.
17. DRAKE, C. A. The Iota function. *J. educ. Res.*, 1940, **34**, 190–198.
18. EDWARDS, A. S. Effects of the loss of one hundred hours of sleep. *Amer. J. Psychol.*, 1941, **54**, 80–91.
19. EELLS, W. C. The California Junior College mental education survey. *Educ. Rec.*, 1930, **11**, 281–291.
20. ELDERTON, E. M., & MOUL, M. On the growth curves of certain characters in women. *Ann. Eugen., Camb.*, 1928, **3**, 277–336.
21. ELLIOTT, M. H. The effect of hunger on variability of performance. *Amer. J. Psychol.*, 1934, **46**, 107–112.
22. ELLIS, H. *Man and woman.* (Rev. Ed.) Boston: Houghton Mifflin, 1929.
23. ESPENSCHADE, ANNA. Motor performance in adolescence. *Monogr. Soc. Res. Child Developm.*, 1940, **5**, No. 24.
24. FLUGEL, J. C. Practice, fatigue and oscillation. *Brit. J. Psychol., Monogr. Suppl.*, 1928, **13**, 1–80.
25. FREEMAN, F. N., & FLORY, C. D. Growth in intellectual ability as measured by repeated tests. *Monogr. Soc. Res. Child Developm.*, 1937, **2**, No. 9.
26. GARRETT, H. E. A developmental theory of intelligence. *Amer. Psychologist*, 1946, **1**, 372–378.
27. GUILFORD, E. E. Essential conditions in the determination of trait variability. *J. appl. Psychol.*, 1939, **23**, 436–439.
28. GILBERT, JEANNE G. Memory loss in senescence. *J. abnorm. soc. Psychol.*, 1941, **36**, 73–86.
29. GOLDFARB, W. *An investigation of reaction time in older adults.* Teachers College, Columbia Univ. Contr. Educ., No. 831. New York: Bureau of Publications, Teachers College, Columbia Univ., 1941.
30. GOODENOUGH, FLORENCE L. Variability of the IQ at successive age-levels. *J. educ. Psychol.*, 1942, **33**, 241–251.
31. GRAY, SUSAN W. The relation of individual variability to intelligence *J. educ. Psychol.*, 1944, **35**, 201–210.
32. GRAY, SUSAN W. The relation of individual variability to emotionality. *J. educ. Psychol.*, 1944, **35**, 274–283.
33. GREENWOOD, M., & BROWN, J. W. A second study of weight, variability and correlation of the human viscera. *Biometrika*, 1913, **9**, 473–485.
34. GUNDLACH, R. H., & KENWAY, G. A method for the determination of olfactory thresholds in humans. *J. exp. Psychol.*, 1939, **24**, 192–201.
35. HARMON, G. E. On the degree of relationship between head measurements and reaction time to sight and sound. *Biometrika*, 1926, **18**, 207–220.
36. HASTINGS. W. *A manual for physical measurements.* Springfield, Mass.: Author, 1902.
37. HENMON, V. A. C., & LIVINGSTONE. W. F. Comparative variability at different ages. *J. educ. Psychol.*, 1922, **13**, 17–29.
38. HERMANS, T. G. The perception of size in binocular, monocular and pinhole vision. *J. exp. Psychol.*, 1940, **27**, 203–207.
39. HERTZMAN, M. The relation of individual variability to general ability as measured by mental tests. *J. educ. Psychol.*, 1936, **27**, 135–144.
40. HILL, A. V. The physiological basis of athletic records. *Report Brit. Ass. Adv. Sci.*, meeting at Southhampton, 1925, 156–173. London: Brit. Ass. Adv. Sci., 1926.

41. HOLLINGWORTH, H. L. *Mental growth and decline.* New York: Appleton, 1927.
42. HULL, C. L. Variability in the amount of different traits possessed by the individual. *J. educ. Psychol.*, 1927, **18**, 97–106.
43. IRELAND, M. W., DAVENPORT, C. B., & LOVE, A. G. *Army anthropology.* In *Medical Department, U. S. Army in the World War*, Vol. 15, Part 1, pp. 100–121, 220–221, 224–225. Washington: Govt. Printing Office, 1921.
44. ISRAELI, N. Variability and central tendency in aesthetic judgments. *J. appl. Psychol.*, 1930, **14**, 137–149.
45. KELLEY, T. L. *The influence of nurture upon native differences.* New York: Macmillan, 1926.
46. KINCAID, MARGARET. A study of individual differences in learning. *Psychol. Rev.*, 1925, **32**, 34–53.
47. KOGA, J., & MORANT, G. M. On the degree of association between reaction times in the different senses. *Biometrika*, 1923, **15**, 346–372.
48. LAMARCK, J. P. B. A. *Philosophie zoologique.* 1809. Reprinted, Paris: Schleicher, 1907.
49. LEUKHART, R. H. The speed of monocular accommodation. *J. exp. Psychol.*, 1939, **25**, 257–270.
50. LINCOLN, E. A. *Sex differences in the growth of American school children.* Baltimore: Warwick & York, 1927.
51. MACDONEL, W. R. A study of the variation and correlation of the human skull with special reference to English crania. *Biometrika*, 1904, **3**, 191–244.
52. MCGEOCH, J. A. The influence of associative value upon the difficulty of nonsense-syllables. *J. genet. Psychol.*, 1930, **37**, 421–426.
53. MCNEMAR, Q., & TERMAN, L. M. Sex differences variational tendency. *Genet. Psychol. Monogr.*, 1936, **18**, 1–66.
54. MANZER, C. W. The effect of fatigue upon variability of output in muscular work. *J. exp. Psychol.*, 1934, **17**, 257–269.
55. MAY, M. A. Predicting academic success. *J. educ. Psychol.*, 1923, **14**, 429–440.
56. MELTZER, H., & BAILOR, E. M. Sex differences in knowledge of psychology before and after the first course. *J. appl. Psychol.*, 1930, **14**, 107–121.
57. MILES, CATHARINE C., & MILES, W. R. The correlation of intelligence scores and chronological age from early to late maturity. *Amer. J. Psychol.*, 1932, **44**, 44–78.
58. MILES, W. R. Performance in relation to age. *U. S. Publ. Hlth. Serv. Suppl.*, 1942, No. 168, 34–42.
59. MONTAGUE, HELEN, & HOLLINGWORTH, LETA S. The comparative variability of the sexes at birth. *Amer. J. Sociol.*, 1914, **20**, 335–370.
60. ODOM, C. L. A study of the mental growth curve with special reference to the results of group intelligence tests. *J. educ. Psychol.*, 1929, **20**, 401–406.
61. OWENS, W. A., JR. Intra-individual differences vs. interindividual differences in motor skills. *Educ. Psychol. Msmt.*, 1942, **2**, 299–314.
62. OWENS, W. A., JR. A note on the effects of practice upon trait differences in motor skills. *J. educ. Psychol.*, 1942, **33**, 144–147.
63. PAULSEN, G. B. A coefficient of trait variability. *Psychol. Bull.*, 1931, **28**, 218–219.
64. PAULSEN, G. B. The reliability and consistency of individual differences in motor control. *J. appl. Psychol.*, 1935, **19**, 29–42, 166–179.
65. PEARL, R. Variation and correlation in brain weight. *Biometrika*, 1905, **4**, 13–104.
66. PEARSON, K. Variation in man and woman. In *The chances of death*, Vol.

1, Ch. 8. London: Arnold, 1897.
67. PETERS, C. C., & VAN VOORHIS, W. R. *Statistical procedures and their mathematical bases.* New York: McGraw-Hill, 1940.
68. PETERSON, H. A., & KUDERNA, H. A. Reliability of school tests of auditory acuity. *J. educ. Psychol.*, 1924, **15**, 145–156.
69. PHILIP, B. R. Studies in high speed continuous work. I. Periodicity. *J. exp. Psychol.*, 1939, **24**, 499–510.
70. PRESSEY, LUELLA C. Sex differences shown by 2544 school children. *J. appl. Psychol.*, 1918, **2**, 323–340.
71. PYLE, W. H. *The examination of school children.* New York: Macmillan, 1913.
72. PYLE, W. H. Sex differences and sex variability in learning capacity. *Sch. & Soc.*, 1924, **19**, 352.
73. REED, H. B. The influence of practice on individual differences. *Sch. & Soc.*, 1931, **34**, 100–102.
74. REED, H. B. The influence of training on changes in variability in achievement. *Psychol. Monogr.*, 1931, **41**, No. 185.
75. RICHEY, H. G. The relation of accelerated, normal and retarded puberty to the height and weight of school children. *Monogr. Soc. Res. Child Developm.*, 1937, **2**, No. 8.
76. ROBSON, G. C., & RICHARDS, O. W. *The variation of animals in nature.* New York: Longmans, Green, 1936.
77. RUGER, H., & STOESSIGER, B. On the growth curves of certain characters in man. *Ann. Eugen., Camb.*, 1927, **2**, 76–110.
78. SCHULTZ, A. H. Studies on the variability of platyrrhine monkeys. *J. Mammal.*, 1926, **7**, 286–305.
79. SCHULTZ, A. H. Variations in man and their evolutionary significance. *Amer. Nat.*, 1926, **60**, 297–323.
80. SCHULTZ, A. H. Age changes and variability in gibbons. *Amer. J. phys. Anthrop.*, 1944, n.s. **2**, 1–129.
81. SHULL, A. F., LARUE, G. R., & RUTHVEN, A. G. *Principles of animal biology.* (3rd Ed.) New York: McGraw-Hill, 1929.
82. SKAGGS, E. B. Sex differences in feeling and emotional disposition in a university population. *J. soc. Psychol.*, 1942, **16**, 21–27.
83. STERN, W. *Differentielle Psychologie.* (Dritte Auflage.) Leipzig: Barth, 1921.
84. TERMAN, L. M. *The measurement of intelligence.* Boston: Houghton Mifflin, 1916.
85. TERMAN, L. M., & MERRILL, MAUD A. *Measuring intelligence.* Boston: Houghton Miffiin, 1937.
86. THORNDIKE, E. L. Empirical studies in the theory of measurements. *Arch. Psychol., N. Y.*, 1907, No. 3.
87. THORNDIKE, E. L. *Educational psychology,* Vol. III. New York: Teachers College, Columbia Univ., 1914.
88. THORNDIKE, E. L., BREGMAN, E. O., COBB, M. V., & WOODYARD, ELLA. *The measurement of intelligence.* New York: Bureau of Publications, Teachers College, Columbia Univ., 1927.
89. THURSTONE, L. L. The absolute zero in intelligence measurement. *Psychol. Rev.* 1928, **35**, 175–197.
90. THURSTONE, L. L., & ACKERSON, L. The mental growth curve for the Binet tests. *J. educ. Psychol.*, 1929, **20**, 569–583.
91. TILTON, J. W. The effect of forgetting on individual differences. *Psychol. Monogr.*, 1936, **47**, 173–185.
92. TODD, T. W., & LINDALA, A. Dimensions of the body: whites and American Negroes of both sexes. *Amer. J. phys. Anthrop.*, 1928, **12**, 35–119.
93. TOUTON, F. C. Sex differences in geometric abilities. *J. educ. Psychol.*, 1924, **15**, 234–247.
94. TRAWICK, M. Trait consistency in

personality. *Arch. Psychol., N Y.*, 1940, No. 248.

95. VERNON, H. M. *Variation in animals and plants.* New York: Holt, 1902.
96. WARDEN, C. J., BROWN H. C., & ROSS, S. A study of individual differences in motion acuity at scotopic levels of illumination. *J. exp. Psychol.*, 1945, **35**, 57–70.
97. WATSON, R. I. An experimental study of the permanence of course material in introductory psychology. *Arch. Psychol., N. Y.*, 1938, No. 225.
98. WECHSLER, D. *The range of human capacities.* Baltimore: Williams & Wilkins, 1935.
99. WEINLAND, J. D. Variability of performance in the curve of work. *Arch. Psychol., N. Y.*, 1927, No. 87.
100. WELLS, F. L. Normal performance in the tapping test. *Amer. J. Psychol.*, 1908, **19**, 437–483.
101. WHIPPLE, G. M. Sex differences in intelligence test scores in the elementary school. *J. educ. Res.*, 1927, **15**, 111–117.
102. WHITING, M. H. On the association of temperature, pulse and respiration with physique and intelligence in criminals. *Biometrika*, 1915, **11**, 1–37.
103. WINSOR, A. L. The relative variability of boys and girls. *J. educ. Psychol.*, 1927, **18**, 327–336.
104. WOODROW, H. Quotidian variability. *Psychol. Rev.*, 1932, **39**, 245–256.
105. WOODROW, H. The effect of practice on groups of different initial ability. *J. educ. Psychol.*, 1938, **29**, 268–278.
106. WOOLLEY, HELEN T. *An experimental study of children.* New York: Macmillan, 1926.
107. YERKES, R. M., BRIDGES, J. W., & HARDWICK, ROSE S. *A point scale for measuring mental ability.* Baltimore: Warwick & York, 1915.
108. YOSHIOKA, J. G., & JONES, H. E. An analysis of children's maze learning in terms of stages of learning. *J. genet. Psychol.*, 1945, **67**, 203–214.
109. YULE, G. U. *An introduction to the theory of statistics.* London: Griffin, 1922.

A DISCUSSION OF SOME CAUSES OF OPERATIONAL FATIGUE IN THE ARMY AIR FORCES*

LESSING A. KAHN
University of Pennsylvania

Definitions

According to the *Flight Surgeon's Handbook*, operational fatigue is defined as a predominately emotional condition found in air crew personnel as a result of abnormal strains being placed on normal individuals.

> . . . flying stress or operational fatigue is used to describe a condition that may be oserved as an *abnormal* flying strain being placed on a normal individual (10).

Its prevalence was marked in members of crews engaged in combat flying. The dangerous and extreme nerve-racking character of combat flying acted as overwhelming stresses upon many so engaged.

The supposition that "flying fatigue" is in any strict sense to be construed as being equivalent to the more general and more subtle syndrome of operational fatigue is, in my opinion, without basis; notwithstanding that such fatigue may enter as a factor contributing to the generation of operational fatigue. Spiegel notes (12) that the term "flying fatigue" as used to describe a clinical state may be manifested by any one of the following symptoms: anxiety, depression, anorexia, dreams, agitation, tremors, loss of confidence, regressiveness, and many others that are psychological in origin; fatigue of this sort is not a reaction to flying *per se*, rather to the conditions under which flying takes place. Hastings (5) makes a most accurate and clear differentiation between these two types of fatigue:

> Flying fatigue . . . means ordinary fatigue and the physical and mental symptoms of it and does not imply that the individual is emotionally sick. Flying fatigue is the same as the fatigue any individual would suffer if he had

* The author wishes to express his appreciation to Prof. Samuel W. Fernberger for reading this paper and offering many helpful suggestions.

(*Editor's Note:* Mr. Kahn served in the Army Air Force as a combat navigator with the rank of 1st Lieutenant. He was stationed in Italy with the 450th Heavy Bombardment Group, 15th Army Air Force, from July until September 1944, when he was shot down in combat over Vienna on September 12th. He was then interned as a prisoner of war in Germany where he had the opportunity to collect case history materials concerning the acute emotional conditions considered in his review.)

insufficient sleep, rest, and relaxation, and had been exposed to the nervous strain of flying.

Operational fatigue . . . is used to describe a typical syndrome of breakdown occurring in essentially stable individuals, who by continued stress, harrowing experiences, and physical fatigue develop an illness which is roughly half fatigue and half emotional illness.

Symptoms

The most common symptoms of operational fatigue lie within the broader context of anxiety states (4)—in which the individual is still attempting to deal with a situation that has long since been eliminated from the present sphere of environment; in short, the individual is attempting to deal with reactions to combat. Moreover, according to Hastings as regards to the direction in which aggression can spread:

> . . . schematically one can hit back at that which threatens one, one can hit at other people, at inanimate things or at oneself . . . [projection, dissociation, introjection, etc.]
>
> In a study made of combat fatigue, etc., 29% of successful combat men state with feeling that they have a personal hatred against the enemy, that they want to kill the men in the enemy fighters—at least when they are in the air. More than 70% state that they develop irritability and quick flaming anger at their crew mates, in a way entirely foreign to their usual feelings and actions, as their operational tour progresses. Such feelings of animosity develop most characteristically during periods of inactivity, when there is little active outlet for their feelings. The monthly dances held at groups bases have been marked by a number of violent fights. For example, on one occasion two squadron commanders, quite close friends, riding back to the post after an evening of moderate drinking, amicably and deliberately decided they "needed a fight," and without any quarrel of any sort, got out of the car and fought violently until one broke a metacarpel bone, after which they amicably climbed back into the car and drove home.
>
> In order to relieve tension, it was not uncommon for men to shoot out the lights with a tommy gun or to shoot one's initials into a wall with a pistol. The feeling was that of just blowing off or smashing something.
>
> Several men have reported, upon close examination, that they have seduced women, not for any sexual gratification, but for the sake of subduing and conquering their defenses.
>
> Enlisted men's barracks are the scenes of consistent and brutal attacks upon the equanimity of new crews coming into them. The old combat man describes to the new men the appearance of a man's brains scattered about a plane by the action of a 20 mm shell, and the like (5, pp. 137–139).

The persistence of anxiety, brought about by flying incidents in which personal security and safety has been menaced, is characterized by a breakdown in adaptive mechanisms. This is manifested by the individual's failure to relax and be free of tension when he is no longer

confronted with the combat situation. Grinker and Spiegel (4) give us a clear picture of the clinical syndrome of anxiety:

> Severe anxiety states result in an intensely striking, unforgettable picture. Terror-stricken, mute, and tremulous, the patient closely resembles those suffering from an acute psychosis. The facial expression may be vacuous or fearful and apprehensive. There are course tremors of the extremities. Speech is usually impossible except for a few stuttering attempts to frame an occasional word. Sudden fits of crying or laughing may occur without reason. Behavior is extremely bizarre, and attitudinizing with apparently senseless gestures, alternates with periods of excessive activity, characterized by running about the ward and leaping over the beds. Terror is one of the principle themes of the patient's behavior. He resembles a frightened inarticulate child, with only a few persistent "islands" of his past well-organized behavior.
>
> Mild anxiety states in contrast to the severe anxiety states present uniform and sometimes monotonously similar clinical pictures. Upon going into battle, anxiety appears in gradually increasing amounts. At first it is kept under control through an effort of will. Under the continued stress of battle, of near escapes from death, of constant exposure to anti-aircraft fire, the anxiety overwhelms him. He develops gross persistent tremor and feels weak, as if his legs would carry him no further. He becomes dizzy so that intelligent thought is impossible.

The picture of the mental state of these men is typical and strikingly uniform as to symptomatology (3, 4, 5, 6, 9, 12). On missions the individual is usually tense, jittery, restless, and inefficient. In the presence of flak or enemy fighter opposition, he becomes overly apprehensive and fearful of the entire combat presentation. He complains that the airplane will be shot down, that he will have to bail out, that the plane will be shot up to such an extent that escape for all will be impossible, or that he will be hit by flak or other missiles. In many cases, free-floating anxiety is experienced along with the usual physiological concomitants of anxiety, such as, dizziness, nausea, feelings of weakness, high blood pressure, rapid pulse, extreme perspiration, and the like (4). More severe symptoms may become evident such as headaches, vomiting, fainting spells, etc. Needless to say, the appearance of these symptoms of fatigue during the course of a mission not only endangers the entire organization, but summarily affects the over-all efficiency of the individual concerned. In the more severe cases, a state of fear and terror develops to the point where bailing out becomes the only possible solution. Levy (7) cites several interesting examples of patients who lay upon the floor of the plane to prevent themselves from giving in to the impulse of bailing out. In between missions, sleep is

disturbed, dreams and nightmares centered about combat experiences, increased tension and irritability, etc., are the rule (**4, 5**).

The appetite is impaired and weight is usually lost (**3**). Excessive drinking and smoking as an attempt to relive anxiety is common; and the need for finding sexual outlets is intensified by the factors of fatigue and anxiety of these individuals (**4, 5**). A frequent complaint (**12**) and at the same time a symptom is the loss of confidence in the ability to perform the job with responsibility, and a conflict occurs with regard to ultimate consequences to the remaining crew members. This brooding results in further conflict and anxiety, which has its most telling effect on an already weakened ego with ever increasing feelings of insecurity and loss of balance.

The following case cited by Grinker and Spiegel illustrates for us the developmental aspects of the fatigue syndrome:

A twenty-four year old airman had demonstrated fine skill as a pilot and was much respected for his aptitude and judgment. During a night mission, the bomber plane in which he was flying co-pilot became lost. The pilot cruised for hours looking for home base, until finally the gas ran out and he instructed the crew to parachute to safety. However, he had misjudged the altitude. Four other members of the crew got out safely, but the patient had considerable difficulty getting out of the pilot's compartment to the escape door. There was much delay and loss of precious time while he struggled to get through; finally by a super-human effort he made it and jumped, leaving the pilot still in the plane. Because of the low altitude, his parachute had just opened when he landed on rough ground, injuring his back. A few minutes later he heard an explosion accompanied by a tremendous flash of flame; obviously the pilot had crashed some distance away. Although he was concerned for the safety of the pilot, nothing could be done that night because of the complete blackness.

The following morning he rounded up the other members of the crew who were safe in the area, and set out to search for the plane and pilot. They found the pilot's body, a charred and smashed mass, among the twisted wreckage of the plane. Clearly, the pilot had not had time to escape before the plane crashed. Eventually the men were picked up by friendly natives and made their way back to base where the patient was hospitalized because of the back injury, which consisted of a fractured transverse vertebral process. In the hospital he at first exhibited marked continuous anxiety, associated with tremor and terrifying dreams, in which he saw himself falling from a plane and crashing on the hard ground. Considerable depression and grief for the dead pilot, who had been a good friend of his, was manifested. The grief was accompanied by some conscious guilt because of the patient's delay in escaping from the plane. This had not actually been his fault; but had it not occurred, the pilot could have jumped in time to save his life. In time the anxiety and depression disappeared to be replaced by a pronounced conversion symptom related to the back injury. The lesion had healed; but the patient continued to complain of

the severe pain. At this point hospitalization was discontinued and the patient returned to his unit in the hope that he would give up the conversion symptom, when back in the environment of his friends and active operational duties. He was given ground duties as an operations officer, and after a short time the conversion symptom disappeared, to be again replaced by a slight anxiety and depression. The patient felt very badly because his friends and class-mates at flying school were participating in missions while he remained on the ground. Accordingly, in spite of the mild anxiety, he asked to be restored to flying status.

After a short period of training, during which he did very well, he was assigned a crew and thereafter participated in seven bombing missions over Tunis and Bizerte, each time going through heavy flak, which severely damaged his plane. He had no increase of anxiety in relation to these missions, and felt confidence in himself. On the eighth mission, however, his tent mate and best friend was shot down in flames over the target. The patient brought his own plane back uneventfully, but after his return he had an intense recurrence of anxiety with tremor. That night he had the first recurrence of the anxiety dream of falling. The next day he continued to have anxiety, but determined to fight it; he said nothing and went on a mission. However, he had so much anxiety that the experience was like a nightmare and he could scarcely keep his mind on the job. In addition, he was harassed by a specific phobic apprehension that the plane was falling off to the right—the direction in which he had seen his friend's plane fall. This phobia persisted, in spite of the evidence of his senses and the instruments that the plane was in level flight. So much did this obsess the patient that he almost crashed the plane on landing, because of miscalculation, greatly upsetting the crew, who reported the situation to the Flight Surgeon. The patient confessed the recurrence of anxiety to the Flight Surgeon, who referred him for psychiatric evaluation, and it was at this point that we first saw the patient.

In the hospital he exhibited moderate anxiety, tremor, and dreams of falling, with a marked phobic response to every aspect of flying. He could not think of planes without strong anxiety. Because he had enjoyed flying, this reaction gave him considerable depression . . . (4, pp. 122–126).

In sum, the crux of the problem of adjustment to the combat situation is the degree of success which the individual is able to maintain in order to cope with the manifold factors effective in the combat situation, namely, threats to the individual's personal safety and security. The initial phases of the fatigue syndrome are constituted of mild anxiety states, in which the individual is attempting to reconcile his continuation in combat and maintaining his own personality intact. When such attempts are uniformly unsuccessful and prolonged, the resulting anxiety states become more intricate, and implications therefrom become more inclusive symptomatically speaking. On the other hand, severe apprehension of the combat situation is the most likely extreme to be manifested by those individuals relatively free from anxiety and other fatigue symptoms.

Causes

Armstrong (1) lists the occupational causes of fatigue as follows:

I. *Physical Agents*
- a. Heat
- b. Cold
- c. Improper clothing
- d. Vibration
- e. Glare
- f. Noise
- g. Wind
- h. Acceleration
- i. Barometric changes

II. *Emotional Stresses*
- a. Physical discomfort
- b. Responsibility
- c. Attention
- d. Concentration
- e. Alertness
- f. Apprehension
- g. Anxiety
- h. Fear

III. *Deficiencies*
- a. Oxygen (anoxemia)

IV. *Toxic Agents*
- a. Carbon Monoxide poison

Not only does combat flying involve the many factors given above, but more comprehensive ones such as geographic and climatic conditions, enemy installations, and opposition in active combat. Among all of the factors contributive to operational fatigue, my experience has shown the emotional stresses to be most potent. Such other causative factors as deficiencies, toxic agents, geographic, and climatic conditions, etc., I have found to be only incidental to the really disposing factor of emotional stress.

Physical Agents

In a consideration of combat stresses which dispose crew members to the sundry forms of operational fatigue, physical factors which operate during a flight play a decisive role. Missions are flown at high altitudes and in close formation to afford a maximum amount of protection from enemy opposition. Oxygen masks, electrically heated suits, steel helmets, and specially built armor suits (flak) are worn on all missions. Pilot and crew members must suffer these physical discomforts for long periods of time and the effects of these precautionary devices are telling. Constant opposition from enemy fighter aircraft, the noises of bursting machine guns, canon, shell hits, flak bursts and hits, the humdrum of engines being throttled back and forth, the rush of the wind, the noise of radio and interphone equipment, and a host of other physical events all contribute to the general emotional and physical tenseness of the situation on hand. The following report by Hastings (5) vividly portrays some of these conditions present on a combat mission:

The airplane, a B-17, on a mission over a distant target in enemy-occupied Europe, had most of its controls shot out by attacking planes before reaching the target, so that the ship was knocked out of formation. The pilot, however,

with the exercise of great skill and strength, persisted in making an effective bomb run. Following this, the lone airplane was attacked by about 100 FW-190 fighters over a period of perhaps three-quarters of an hour, during which time extraordinary damage was done to the plane and crew. Virtually all of the crew were wounded, three severely, and one became anoxic as a result of the simultaneous explosion of a 20 mm. cannon shell next to him, and the severance of his oxygen system. Almost all of the control cables were cut in various places, the oxygen, hydraulic, and electrical systems were knocked out, the interphone and radio systems destroyed, a small fire started in the bomb-bay, large holes were put through both wings, holes were in both propellers, bomb-bay, fuselage and nose; the tail assembly received so many direct cannon hits that it vibrated violently, and, after inspection by the flight engineer, was expected to tear off entirely at any moment.

We must be cautious, however, not to localize any of these stresses and claim them to be the only causative factors operating to produce fatigue. Thus, in discussing physical agents as participating in the production of operational fatigue, we shall also note that each and every one listed is highly charged with emotionally disturbing aspects.

During the first five missions most crews would have encountered the harrowing experiences which were the normal events for heavy bombers operating from this theatre [Eighth Air Force stationed in England]. Watching close in and constant enemy fighter attacks, flying through seemingly inpenetratable walls of flak, seeing neighboring planes go down out of control, and at times explode in mid-air, returning with dead . . . (5).

High altitude flying in itself constituted one of the greatest hazards of combat flying—the most important aspect of which was the constant need for oxygen in proper amounts in order to function at maximum efficiency. The intense cold and the danger of frost bite were constant reminders to all crew members of the latent dangers at hand.

My first two or three missions bring to mind the extreme uncomfort of high altitude flying. Equipped with inadequate heated suits, I would lie huddled in the forward end of the nose of the ship, where I could at least observe the instrument panel and record with frozen fingers the progress of the plane. I frankly confess that I was completely oblivious of all that was going around me, that is, fighter and flak opposition. Under these conditions, I managed to perform my duties, although at a minimum of proficiency.

Glare played an important part to offend crew members, especially pilots and gunners who had to be at all times at top operational efficiency regardless of such disturbing factors. The great feeling of mutual dependence which crew members held for one another often caused them to endure the most trying of hardships, such as flying into the sun and holding such a formation position until the tactical situation changed. Gunners constantly had to scan the skies and be on the

alert for oncoming fighters—enemy fighters that purposely held the sun behind them in attack so as to hamper our own gunners.

Whatever the rationale operative in these physical conditioning factors, one point is outstanding, namely, while they may not in themselves be sufficient to bring about a partial or complete personality breakdown, yet their effect was quite apparent.

Geographic and Climatic Conditions

Morale and general fitness of crew personnel, preparatory and subsequent to combat duty, played a vital role in maintaining top efficiency. The need for stimulating and diverting environmental changes was recognized early in the war, mainly due to the many combat reports which stressed such factors as positive measures in the prevention of fatigue and other mental disorders.

Tropical or sub-tropical climates with their heat, humidity, and frequent rainfall were novel to fliers, and had definite effects upon them —producing run-down physical conditions, loggy feelings, and lazy attitudes (3). Semi-civilized countries, such as was the case in the South Pacific area, that lacked even a semblance of modern sanitary facilities, constituted a constant menace to the health of the men. Because of geographical location and difficult supply lines, food and nutrition in general became another problem in keeping the morale and general fitness of the men at a high level. Repetitous diets with little or no variation acted to make men disgruntled and malcontent, and made for the development of unwholesome attitudes with respect to their duties. Dougherty reports:

> During the month of November 1942 physical fitness in the pilots underwent a rapid decline. There was an acute shortage of fresh fruits, meats, and vegetables, and the foodstuffs that were available contained many gas forming foods. Combat missions had to be flown in the morning and in the afternoon, and for a period of three weeks the chief food for the noon meal was a stew which was highly seasoned and which contained considerable gas forming elements. As a result there was a great increase in minor gastro-intestinal complaints such as heart-burn, gastric-distress, and at times nausea. The constant repetition of certain foods caused a distinct distaste for them and a very definite loss of appetite, and the food was not eaten. Consequently a diet which was satisfactorily balanced became one which did not meet the nutritional demands, and malnutrition characterized by weight loss developed. Vitamins were available in the foods and in the supplemental multi-vitamin capsules; however, vitamin C was inadequate and remained so throughout the period. Fortunately, some fresh food supplies arrived in the early part of December, and the physical condition of the men improved (3, pp. 36–37).

In the European theatre of operations, the over-all situation was not

radically different from that in the South Pacific. For example, I found sanitation facilities in South Italy (from which a greater proportion of combat missions were flown) were none too good. Veneral diseases were high because of the highly infected population living close by to operational fields, and recreational facilities in near-by towns were in in many instances limited due to the war-torn state of affairs. Toward the end of the war, the pace of battle progressed so rapidly that men were not certain about where they were going to sleep from night to night. The constant demands of battle tactics kept most flight personnel constantly on the move without much opportunity for recreation of a too varied sort.

In short, factors which were in many instances completely out of the control of army commands often contributed towards making the lives of combat men monotonous and void of any diversion from the nerve-wracking business of combat. This lack often, while not acting as the precipitating cause of operational fatigue, certainly contributed important aspects to its development and fruition.

Emotional Stresses

Probably the most important aspects of the personality structure of combat personnel were those to which we commonly refer as emotional. In any given combat situation, physical and environmental stress void of any emotional contributing elements is without meaning. The effect of enemy gun fire, for example, produced considerable tension and is in one sense to be looked upon as a physical stress; yet in a stricter sense the emotional "charge" which such a condition produced was more manifest and effective. Among these conditions the following are important:

1. Enemy aircraft may be encountered and may result in seriously wounded or killed crew members; or "bailing out" or crash landings in enemy territory; flying over mountains, over jungles, and over water.
2. Geographic and climatic conditions, such as missions through storms, overcasts, and all sorts of altitude conditions that conjure up all sorts of possibilities as to the final outcome of the mission (for example, getting lost, crash landing, running out of fuel, parachuting, and the like).
3. Base conditions, that is, location of bases with reference to enemy operations; how frequently they are subject to attack by enemy bombers, strafing raids, and the like.
4. And finally, each and every member of a crew feels responsibility toward his country and the principles for which he is fighting; that is, the men feel the necessity of accomplishing a mission not because they are given credit on the "board" for it but because they see it as a stab at the enemy. Their participation means that many more bombs. Members of a crew hold themselves

responsible for the fear they manifest, not so much of the enemy fighters or flak that they undoubtably encounter as of what their buddies and squadron organization will think of them.

The requisites of a battle situation suggesting any of these aforementioned possibilities produce extreme emotional stress in all concerned. These are the problems which properly set the stage for emotional conflicts—conditions and situations in which the factors of environment, physiological and emotional stresses, and so on are mutually reinforcing.

Responsibility. The responsibility which command members of a flight crew feel for subordinates is probably one of the most potent motivating factors operative in personal relations. Does a pilot, for example, consider himself adequately responsible for the safety and well being of his crew? On non-operational flight duty, there is little emotional stress on the part of a pilot in this one respect. There is actually little need for it, since such missions do not call upon the pilot to exercise anything above ordinary operational control over the members of his crew. However, in battle, the situation is quite different. In addition to routine flight command responsibilities, he must also display judgment about the battle situation, he must effect decisions of paramount importance, bearing in mind at all times his responsibility to the command and its mission (of which he is a vital part) and to his crew members. At times, the emotional burden which this responsibility imposes is exceedingly heavy, and there are many instances of pilots (as well as other members of a crew) breaking under these demands. Wrong judgments are feared most of all, for the decision formulated in battle must be quick and accurate if the crew is to survive. Supposedly the selection of pilots for combat has been determined by a man's capability of exercising such responsibility without serious psychiatric consequences. However, this selection is not always effective for as we have mentioned earlier, it often happens that apparently stable individuals break unexpectedly with disastrous consequences for all concerned.

I recall from my own experience several first pilots who resigned their position to accept a co-pilot's position because, as they claimed, they did not wish to assume the responsibility of nine crew members. They felt that they did not possess the necessary requisites for making quick decisions, for evaluating combat situations, and the like, contending that a wrong move on their part would bring serious consequences for which they did not want to feel everlastingly responsible.

I also recall in connection with this discussion the incident involving a B-24 pilot who had lost a good portion of his crew by the exercise of a wrong

judgment, namely, he pulled his ship out of formation with 3 engines in good condition, and thereby became an easy target for enemy fighters.

The question of responsibility does not apply to the pilot alone, but extends to all members of a flight crew. The navigator is responsible to his crew to see that they reach their target and that they have a safe return route. He must be cautious with regard to his computations, the accuracy of the ship's course, and many other details for which he is solely accountable. Mistakes on his part may not only affect the men in his own ship, but in many cases (e.g., when he is in the lead ship of a bomber formation) he is performing for a great number of other ships. One can readily realize the tremendous emotional stress which a navigator must undergo who is entrusted with these responsibilities.

A navigator, during the heat of an intense fighter attack, abandoned his position and in great fear and anxiety fell to the floor of the plane for protection. The plane was crippled as a result of the attack and it became separated from the rest of the bomber formation. The navigator, by his action, had failed to keep track of the bomber's course and it turned out subsequently that the pilot brought the plane over a flak area and was shot down. The navigator's neglect to perform his duty can be attributed not only to his fear and anxiety, the resultants of the battle situation, but to a definite lack of responsibility toward his crew mates. The literature is replete with cases in which men have been overcome with fear and anxiety, but who have refused to give up their position because of the responsibility that they had toward their buddies.

The bombardier is held accountable for dropping his bombs on the target, not only that the mission might be considered successful, but that repeat missions will not become necessary and thereby further endanger a large force of men and planes. Bombardiers, because of the stresses of combat conditions, become emotionally upset, so much so, that repeat runs on the target become necessary-- this entails one or more additional runs on the target that may very well be heavily defended by flak. Anxiety and fear, under these circumstances, are magnified considerably in all personnel concerned. Bombardiers who displayed these characteristics soon became known and stood out in the squadron; their unfavorable reputation often caused them to become shy and seclusive; to complete the picture, the typical symptoms of a war neurosis became evident.

On a bombing mission directed at a rail-bridgehead located in northern Italy, I recall the following details pertinent to our discussion. This was a very heavily defended target, since it was a main line of supply for the German forces in Italy. The task was to bomb the bridge and thereby hamper the supply flow. The group bombardier had had considerable experience in combat. It turned out that three bomb runs were made on this target; losses were much higher than necessary. The feeling of tension release which one experiences

after making a bomb run (in getting out of a flak area) cannot be minimized. In repeating the bomb run twice, with the enemy's flak becoming more and more accurate on each succeeding trial, I venture to say that not a small number of men would have gladly shot the lead bombardier.

Gunners in all positions were held responsible for defending the ship. This task was very important for interference from enemy aircraft could and often did cause a mission to fail. Gunners, therefore, held a unique position in preventing harm from coming to their buddies and were indirectly responsible for the success or failure of the mission.

The following details of a mission were reported to me. Returning from a mission, it was necessary for the group to cross the Adriatic Sea in order to reach their base of operations in Italy. It was standing operating procedure that upon reaching the sea, all gunners would be relieved, since there was then little danger of attack by the enemy. It was therefore customary to raise the ball gunner. However, on this particular occasion, for some unknown reason, the other gunners whose responsibility it had been to raise and release the ball gunner failed to do so. The ship ran out of fuel and it had to be ditched. It was too late to do anything about the trapped gunner.

A second case concerns a nose gunner in a B-24. It was customary upon approach to the bomb run area for each crew member to don his flak suit. In the case of pilots, nose, and ball gunners, they are helped into their suits by nearby crew members. In the case of the nose gunner, it is the responsibility of the navigator to do so. Upon reaching the target area, the gunner began screaming over the interphone for his flak suit. The navigator was busy charting the bomber's course. It was the navigator's feeling that at the time he was in no position to give attention to anything other than his job. An 88 mm. shell burst just ahead of the ship and a piece of flak killed the gunner.

In the case of the fighter pilot, one would be inclined offhand to think that they had little responsibility to exercise—perhaps, only in the sense of completing a strafing mission, or a dive bombing mission, or just escort for a group of heavy bombers. Upon closer examination, however, it will be found that their responsibilities required as much of them as was the case with bombardment personnel.

Tactical expediency in the case of fighter formations necessitated the use of two-plane elements. There was a lead plane and a wing ship. In this manner, it was difficult for an enemy plane to sneak up behind one without the other being aware of it. In combat, the lead plane did most of the aggressive fighting and the wing plane afforded protection. Credit for destroying enemy aircraft went to the lead ship by virtue of its function. Honors in battle were strongly vied for and intense jealousies were evident between lead pilots and wing pilots, since it was held that undue credit was being given to the lead pilot. The need, however, for this arrangement of planes in combat was manifestly neces-

sary for the safety of all concerned and the necessity for having a responsible wing pilot was very apparent. Unfortunately for both pilots, the wing pilot often forgot this responsibility and sought after a little glory of his own, with the result that both planes would suffer the almost inevitable consequences.

A second responsibility of fighter pilots was manifest on cover or protective missions, that is, in escorting heavy bombers and protecting them from enemy fighter opposition. The need for this cover was essential and bomber personnel were always very grateful because of it. Failure on the part of fighters in meeting the bombers at the proper time led to serious consequences. A bomber could not weave about the sky in wait for its cover, since fuel consumption was a vital problem in every attack and target time was strategically set. The result of any failure of fighters and bombers to coordinate was evidenced by the high bomber losses incurred on the raid.

Fighter pilots were always attracted to grounded enemy planes and other land targets. Their failure to meet pre-arranged schedules with bombers was often due to this fact; taking time to strafe grounded craft, supply trains, and other targets delayed their time to such an extent that for all practical purposes they were useless.

In sum, responsibility as an emotional stress entails a number of factors. It consists primarily of an alertness, apprehension, and attention to the many details that go to make up the job that has to be performed. It is not just a realization that so much has to be accomplished; rather it is a knowledge of the necessity of performing that job. It consists of a concentration of energies and attention upon the ramifying points entering into performance of duty as a whole. Failure to achieve any of these components ultimately leads to a conscious disavowal of responsibility or unconsciously to gross mistakes that in the final analysis must be interpreted as failure to act responsibly. The end product of all of this is, of course, some sort of mental conflict.

The concept of responsibility may be looked upon from another point of view, namely, lack of responsibility as a function of or symptom of an already distorted personality. Thus, as was the case many times, crew members who had already acquired conflicts of one sort or another, such as, anxiety states, fears, and the like were by reason of their condition unable to act in a responsible manner. They were not only a danger and detriment to themselves and their crew mates, but in a larger sense to the entire formation of which they were a small part.

A pilot failed to pull out of formation after being hit by a flak burst. It was standing operating procedure that when a ship was seriously hit and there

was some danger of explosion, that the pilot was obliged to get the ship out of the formation in which it was flying. This pilot neglected to do so; the result was that he blew up and caused a number of other ships to go down with him. Upon subsequent investigation, the following facts concerning the pilot were brought to light: he had somewhat over 35 missions; he had been of late speaking of his extreme fear and anxiety of being hit by flak and catching on fire; he had exaggerated anxiety symptoms before being briefed, and had expressed the hope before briefing time that the mission for the day would be an easy one; and the like. There is little doubt that this pilot was mentally incompetent at the time he was flying the mission.

The weight of responsibility, its conditions and demands was at times so highly charged that emotional stability was made impossible. On the other hand, responsible behavior could not be exacted from some individuals since they had already succumbed to some form of personality disorder.

Fear. It was frequently asserted by combat personnel that there was no one man who at some time or another did not experience the harsh feeling of fear before or during a mission. The truth of this statement becomes apparent when one takes into account the many manifold experiences to which the combat man was constantly subjected. There is then little wonder that we list fear as one of the major stresses operating at all times to distort and disorder the personalities of combat men. It is not so much the fact that fear operates independently of any other stresses already mentioned or to be mentioned; rather it is the unique manner in which it manifests itself during the daily routine of flight personnel which constitutes fear a special type of problem. The common experiences of flak, gun fire, explosions, and the like were in themselves great dangers and were capable of producing fear reactions; however, in a great number of instances, anticipatory fears of situations produced more marked reactions (5). Perhaps it was the mystery connected with these possible episodes (e.g., bailing out, ditching, being strafed, etc.) which made them all the more fearful.

The development of fear follows a definite course; a course which parallels the combat career of every person so engaged. In a number of cases, persons responded to fear in new and resourceful ways and eventually overcame it; in other instances, it was fear that won out. Hastings describes the development of behavior in which fear plays a most important part:

1. On arrival at an operational station the men were first of all insecure and defensive, both consciously and unconsciously, and this was apparent in many ways. In action they were either overly self-assured, or particularly diffident, usually the former. In speech they might be either loud and continuous, or in a few cases "mouse-quiet." They either spoke continuously of combat or

avoided it completely. They either ridiculed and paid no heed to any advice by experienced men or they took in every possible word of it. They tended at first to drink more than the others. They did not accept the possibility that they would ever be afraid, and openly spoke of the older men who mentioned fear as being "flak-happy" or spiritless. It was quite easy to spot a new group of officers at a table in the mess or in a group in the lounge, even if one knew none of the personnel of the station.

2. These defensive attitudes and mechanisms began to disappear quite quickly after the men had four or five raids, and had seen and felt the real factors in the combat situation. This process was frequently conscious to a large degree, and in many cases, the men spoke of the change in themselves, and shamefacedly deprecated their former cocky attitude. At about this point it was frequent for them to "over-swing," and to be very conscious of their anxiety, having somatic symptoms, and paying attention to them, and feeling quite hopeless about their chances of survival, sometimes in consequence, getting careless of technique, equipment and the like.

3. The third stage of evolution took place at roughly the tenth raid, by which time one or more of several factors had helped effect a further change: the man had experienced fear and by now knew that he could deal with it; he found that care and skill and coolness in the pilot and crew had a real bearing upon the question of his return; he saw that his crew and his airplane could withstand catastrophe; he developed an "esprit de corps" in regard to his squadron, and was now really part of it. He developed for the first time a sense of his responsibility to his mates, and to the formation. At this stage which continued sometimes until the end of his tour, the men were effective, careful, fighting men, quiet and cool on the ground and in the air. They attained a sort of tranquility in spite of their anxiety. They had very little need for defensive mechanisms of any sort to deceive themselves or anyone else. They talked easily and quietly, drank little except on pass, and expended virtually all of their attention and interest on the job. When they did go on pass and over-indulged they usually did so in a peculiarly deliberate way, believing over-indulgence was a cathartic sort of release of feelings, which they felt to be useful. They were drained of most feelings other than those having to do with combat. No values existed other than those meaningful in combat.

4. Frequently a fourth stage of evolution gradually took place by the beginning of the last five raids. Its components were probably to a large extent physiological, the results of continuous prolonged fatigue and fear. It consisted, in its most extensive form, of a state of insomnia, fatiguability, weight loss, anorexia, indifference, restlessness, loss of concentration and interest and efficiency, marked irritability, loss of libido, and a fairly marked depression with retardation. It did not necessarily include all of these components, but perhaps only several of them; these symptoms could not be considered in the category of neurotic manifestations, in that they did not cloud the real issue from the flier's consciousness. He had usually complete insight into their cause and mechanism. They also did not give him any "secondary gain," in that he did not accept the release from combat that they might have afforded (5, pp. 20–25).

The element of fear may be the result of any of the many factors

encountered in combat flying. Some feared take-offs, for in many types of combat ships the stability of the craft decreases with bomb load and increased use in battle. Stories became numerous of accidents which occurred on take-off and the results of these accidents often acted to create phobias in inexperienced pilots. An engine may fail on take-off, or a tire may blow out as the plane rushes down the runway—with a full load of gasoline and thousands of pounds of bombs such incidents became tragic affairs. And those who have witnessed such happenings readily know the mask of fear on the faces of crew members as they rushed from the burning plane, which, in a matter of minutes, will be blown to bits by the exploding gasoline and bombs (2).

Night missions were especially feared by most combat men. The uncertainty of night flying and its added hazards made for extreme tension and anxiety. Such conveniences (considered in the United States as common necessities) as lighted runways, guide markers, formation lights, radio aids, etc. were not routinely employed in combat for reasons of security. This lack made night take-offs in flying extremely hazardous, and pilots as well as other crew members reacted accordingly with signs of fear and anxiety. The proportion of accidents between daytime and nighttime flying was consistently higher in the night missions.

On daytime raids, the fear of take-off is just as intense as it is on night missions. Perhaps the most critical moment for all crew members simultaneously was that of the take-off. One is sure of a certain degree of success in connection with the mission as a whole once the ship becomes air-borne. Erroneous as this feeling may be, the psychological boost which occurs in combat men once they have become air-borne is not to be minimized. Watching a take-off mishap impresses one with the extent of helplessness manifested by the crew members involved. In a sense, this is not true of the many other dangers connected with flying, wherein certain defensive measures can be employed. The altitude attained on take-off does not allow for freedom of maneuver or the use of adequate safety measures, and when an accident occurs the end is inevitable: ships are completely blown up or consumed by fire in the space of a few minutes. The observer is impressed with all of this, and in some respects there is a certain amount of comfort in the rapidity of it all. Many combat men become fatalistic after a short time in combat, and experiences of such accidents seem to offer some satisfaction in the hope that their end, if it should come, might be a similar one.

That take-off and landing accidents were traumatic situations for the production of fear reactions certainly is undeniable and there can

be little doubt that almost every combat man at one time or another witnessed such occurrences. This being so, the extent to which these experiences acted to produce fear and anxiety cannot be overemphasized. But, as we have already mentioned, combat experience along with the general personality configuration acted to minimize or to emphasize the final effect produced.

Fear of flak and fighters formed two other bases for fear of a most effective sort. At the beginning of a tour of duty, the fear of flak was not very apparent; however, when flak came into close range and combat personnel actually tasted some of its "bitter fruit," then fear of it became in many instances obsessional. The inexperienced flier did not for a moment realize the nature and complete effectiveness of this type of weapon, and for this reason new combat men tended to underestimate the destructiveness of flak. The more common notions about flak were: to be effective, flak had to burst very close to a plane; that enemy flak batteries were not sufficiently accurate to score direct hits; and that there were not enough flak batteries around any single target area to be cause for alarm. Experience, however, very rapidly changed such opinions. It was learned that flak, like shrapnel, was very deadly; that its range of effectiveness was wide: that the Germans had their guns radar controlled and that a fair degree of accuracy was possible; and that the number of guns around many targets ran into the hundreds. It was not at all surprising to note the extent of change which did take place after contact with this weapon.

It was common practice for new crews to complete their first few missions over relatively easy targets, that is, targets which were not heavily defended and at which little fighter opposition could be expected. The rationale of this procedure was to bring about a gradual introduction to the rigors of combat. As the missions became "tougher," one could notice a decided change in the behavior of the men: they became restless and revealed considerable anxiety about the type and objective of their next mission; anticipatory fears developed connected with their position in the formation which connoted varying degrees of vulnerability by flak and fighters; and, in brief, a great many other symptomatic changes of personality attendant upon fear and anxiety. There were many times in my own experience when flak was so heavy over a target that the expression "you could roller skate on it" was a somewhat accurate description. Such targets as Vienna, Wiener-Neustadt, Munich, Regensburg, etc. were reputed to have had many hundreds of flak guns for their defense.

The accuracy which flak achieved was in many instances outstanding. Guns were radar controlled and range estimation was thereby easily and accurately computed. With any large formation of bombers, in the time that it took two or three groups (a group consisted of approximately twenty-five planes) to pass over the target, flak was very accurate and deadly. Thus, when formation notices were posted, the men would be very anxious to know in what part

of the formation they were to fly. As a matter of preventive psychiatry, these formation notices were often restricted to the time of briefing, so as not to cause undue anxiety among the men who were to fly in the rear echelons.

To repeat, first contact with flak usually did not arouse many fears in combat men, for inexperience with the effects of flak, wrong notions concerning its actual and potential danger, and the practice of starting new men on relatively easy missions were operative in the early stages of a combat career. With every mission, however, crews learned what flak could do: physical injury by flak was very painful and many times fatal; damage to planes occurred in places which one would never imagine to be vulnerable; and most effective was explosion due to a direct hit in the bomb-bays which set off the bombs or gas tanks. Combat men witnessing mid-air explosions had their fears enhanced considerably by such incidents.

An enlisted man who had completed fifty-two missions had an extreme fear of flak, the mere sight of which caused him to be paralyzed with fear. Often he would "freeze" on the toggle bombing switch and on several occasions he would not close his bomb-bay doors until called by the pilot to do so. While accumulating his fifty-two missions he had witnessed eleven of our aircraft shot down; with the eleventh craft were lost two of his close friends. On the day he was grounded and hospitalized his own crew was lost to flak—exploded over the target. After hearing of this, he expressed an ardent appeal to discontinue flying altogether. (From a report of a Flight Surgeon attached to a B-26 Bomber Group).

In sharp contrast to the reaction to flak was the effect of fighter opposition in the production of fear reactions. Fear of fighters was not a common experience among crew members. As it was often put, "At least one was able to do something about the damn things;" this was not exactly the case with flak. Those men who did react with fear and anxiety toward fighters did so because of the unfortunate and disastrous traumatic experiences with them. Formation flying was not only designed to bring about optimum conditions for bombing a target, but also to maintain a maximum of protection against fighter opposition. The possibility of becoming disabled for any one of a great number of reasons and having to leave the protection of the formation loomed large as a potent fear producing stress. But the factor of self-defense in being able to fight back gave one a chance to alleviate his fears; the personal implications of the threatened danger were minimized through a constructive attack upon the danger itself.

Toward the closing phases of the war our fighter planes were so numerous and the protection so efficient, that enemy opposition directed against our bombers became inconsequential. As a result of our superi-

ority personal dangers in this direction were considerably lessened, and the incidence of fear reactions correspondingly decreased.

Anxiety. Aero-anxiety ranks next in importance to the factor of fear in the production of operational fatigue (8). To repeat, the incidence of any of these causal agents often went hand-in-hand and this is especially the case with anxiety and fear. Many times it was difficult to differentiate between fear and anxiety; it was very rare that one occurred without the other. This is understandable in view of the nature and consequences of fear itself. Thus, one of the after-effects of fear is anxiety—as manifested in anticipating the same incidents which originally produced the fear. With an intensification of the anxiety state there is a corresponding increase of fear and a lessening of responsibility. In a great measure, all the factors stressed as important in the production of loss of responsibility and of fear are likewise operative in the production of anxieties.

Summary

We have attempted to make clear some of the more important agents influencing operational fatigue—defined as abnormal flying strain being placed on a normal individual. We have seen that they are very complex; that they do not occur in isolation, but rather tend to interact and, in a number of instances, are all present together. In this respect, each becomes a function of every other, and the occurrence of one becomes a condition of the eventual appearance of the others. Thus the burden of supposedly excessive responsibilities brought on the retardation of performance, and prepared for the emergence of fear and anxiety. In an analogous manner, experiences which produced fear also brought with them anxieties and loss of responsibility.

The question of predisposition toward personality breakdown has not been treated in this paper for the obvious reason that at the outset *the assumption was made that all personnel engaged in the hazardous business of flying could succumb to operational fatigue.* It is debatable, of course, whether environmental, climatic, emotional, and similar factors are causative in a true sense, or whether they are the triggers that set off neurotic behavior in a predisposed personality. Clearly, the literature indicates that those men who did break under the strain of battle would in a majority of cases not have done so under the conditions of peace-time living. The predispositions which they might have possessed would have been of little significance in civilian life; however, because of the extreme and prolonged types of experiences in combat, they were bound to react in violent ways. No amount of psychiatric screening could have eliminated these individuals for their condition was in no

sense deep-seated enough to allow for adequate and complete airing. Rehm (**9**) concludes this matter nicely:

It is a matter of general understanding and acceptance by medical personnel and laymen that air crewmen are a highly specialized and trained group. In recognition of the severe strain to which these men are exposed, diligent care has to be taken to see that they are maintained in the finest physical condition. However, other factors, less tangible, which have a direct bearing upon the ultimate combat efficiency of these men, have been placed more or less in the background. . . . These factors are the psychological disturbances which arise in combat flying personnel—the "gremlins" of the mind, in whose grip even the strongest willed man is powerless.

BIBLIOGRAPHY

1. ARMSTRONG, H. G. *The principles and practice of aviation medicine.* Baltimore: Williams and Wilkins, 1943.
2. CONRAD, C. D. The combat man presents himself. In D. G. Wright (Ed.), *Observations on combat flying personnel.* New York: Josiah Macy, Jr. Foundation, 1945.
3. DOUGHERTY, J. E. Flying fatigue—The effects of four months combat flying in a tropical combat zone on fighter pilots. In D. G. Wright (Ed.), *Observations on combat flying personnel.* New York: Josiah Macy, Jr. Foundation, 1945.
4. GRINKER, R. R., & SPIEGEL, J. P. *War neuroses in North Africa.* New York: Josiah Macy, Jr. Foundation, 1943.
5. HASTINGS, D. W., WRIGHT, D. G., & GLUECK, B. C. *Psychiatric experiences of the Eighth Air Force.* New York: Josiah Macy, Jr. Foundation, 1944.
6. KAPLAN, A. J. Emotional disorders of pilots in Assam, India. In D. G. Wright (Ed.), *Observations on combat flying personnel.* New York: Josiah Macy, Jr. Foundation, 1945.
7. LEVY, N. A. *Personality disturbances in combat fliers.* New York: Josiah Macy, Jr. Foundation, 1945.
8. LAYDEN, M. Experiences with anxiety states in combat flying personnel. In D. G. Wright (Ed.), *Observations on combat flying personnel.* New York: Josiah Macy, Jr. Foundation, 1945.
9. REHM, R. Fifty missions over Europe. In D. G. Wright (Ed.), *Observations on combat flying personnel.* New York: Josiah Macy, Jr. Foundation. 1945.
10. REINARTZ, E. Psychiatry in aviation. In F. J. Sladen (Ed.), *Psychiatry and the war.* Baltimore: Thomas & Co., 1943.
11. SMITH, T. The emotionally unstable. In D. G. Wright (Ed.), *Observations on combat flying personnel.* New York: Josiah Macy, Jr. Foundation, 1945.
12. SPIEGEL, J. P. Effects of combat flying stress. In D. G. Wright (Ed.), *Observations on combat flying personnel.* New York: Josiah Macy, Jr., Foundation, 1945.

A REVIEW OF LEADERSHIP STUDIES WITH PARTICULAR REFERENCE TO MILITARY PROBLEMS

WILLIAM O. JENKINS
Indiana University

INTRODUCTION

The present report summarizes and reviews selected references from the available literature dealing with the problem of the selection of leaders in various fields. The primary interest in preparing the article was to provide a summary of techniques and results that would be of value to psychologists dealing with problems of selecting leaders, particularly in the military field.* For this reason, no attempt has been made to cover the extensive literature concerning dominance and "leadership" with other than human subjects.

The primary factor considered in selecting items for inclusion in the present article, in addition to their relevance to military selection problems, was whether or not the material was empirical in nature. For illustrative purposes a few speculative reports have been included, but the main emphasis is on the presentation of research findings.

No attempt is made to treat the theories of leadership which have been proposed. Few of these theoretical expositions have presented hypotheses concerning the various aspects of leadership which are testable. Furthermore, none of them has been comprehensive and systematic to the extent of accounting for the obtained information. In a late section of this article certain general principles and hypotheses extracted from the available literature are presented.

For the purposes of this article the dictionary definition of leadership as the act of guiding or directing the behavior of one or more individuals may be employed. A more adequate operational definition should derive from future research on this problem.

For clarity of presentation the various studies of leadership have been divided into five groups: (1) industrial and governmental investigations, including studies of executives, administrators, supervisors, foremen, etc.; (2) studies of scientific and professional personnel; (3) in-

* This article was originally published as No. 190 in the AAF Aviation Psychology Abstract Series issued by the Psychological Branch, Office of the Air Surgeon, Headquarters Army Air Forces, Washington, D. C. The date of publication was Sept. 20, 1945. The purpose of the original report was to present typical procedures and results obtained as background information for Army Air Forces Aviation Psychologists working on problems of leadership.

vestigations of the activities of children in pre-school and extra-school situations; (4) studies in the school situation; (5) military leadership.

Industrial And Governmental Investigations

Two of the earliest studies of the characteristics of business executives in this field were published by Gowin in 1915 and 1918 (**26, 27**). In the first study a questionnaire was submitted to approximately 1,000 executives, 225 lesser executives, and 200 professional personnel. Data were also gathered on 222,000 insurance policy holders. Information concerning various characteristics of these individuals was requested. Measures of statistical significance were not presented, but differences between the major executives and the insurance policy holders were reported with regard to height and weight, favoring the former group, and the executives were reported to have been subjected to somewhat stricter selection than the professionals as indicated by a lower coefficient of variation. The executives did not differ greatly from the other groups with regard to age at marriage, number of offspring, and similar items. "Attitude toward life work" was rated by the investigator on a ten-point scale from ten (continued one line of activity throughout life) to one (changed line of activity three or more times). The results for approximately 200 business executives were compared with a control group of "professionals." No outstanding differences appeared.

In Gowin's second study 276 business executives were asked to rank the importance of a number of qualities for administrative ability. The qualities ranked at the top were judgment, initiative, and integrity; those ranked lowest were refinement, appearance, and sense of humor. Laird's (**33**) findings were similar to these. No definitions of these concepts were presented nor was further use of the data reported.

A number of other studies involved the administration of questionnaires to business executives. In one study, Taussig and Joslyn (**55**) investigated the social classes from which American business leaders are recruited; attempted to determine the proportionate contribution of each social class to the supply of business leaders; and proposed to study the relative influence of heredity and environment on such disparities as might exist between the representation of the classes among business leaders and their representation in the population at large. A sample of 15,000 business leaders was selected from a register of directors and complete returns were received from somewhat over 7,000 cases on the nine-item questionnaire which was employed. The modal individual emerging from the reports of this study may be described as follows:

president (for less than ten years) of a manufacturing or mining organization, the gross income of which is between one million and five million dollars per annum; a background of living in a community with a population over 500,000 in New York State; between the age of 51 and 52 at the time of questioning and between 41 and 43 at the time of entering the business; a college graduate with no formal business training who reports little or no help in starting in business.

The authors conclude that the results suggest that lack of native ability plays a greater role than lack of opportunity in the failure of the lower occupational classes to be as well represented in the sample as the higher classes. The evidence, however, was negative rather than positive in nature.

Starch (**54**) sent a questionnaire to fifty executives making salaries over $50,000, to 50 making between $7,000 and $50,000, and to 50 with salaries below $5,000. In the opinion of the members of these three groups, four characteristics of executives were important: ability to think, inner drive, capacity to assume responsibility, and ability to handle people. The proportion of the high salary group rating ability to think as important was 72%; the percentage of the middle salary group considering this aspect important was 58%; and the percentage of the low salary group was 46%. For the other characteristics the percentages, respectively, were: inner drive 58%, 42%, and 20%; capacity to assume responsibility 68%, 54%, and 20%; and ability to handle people 84%, 72%, and 86%. On the basis of these results Starch proposed a formula to account for differences in earning potential where "executive achievement" was the product of "drive" and the sum of the factors of "ability to think," "capacity to assume responsibility," and the "ability to handle people." Only the results of the questionnaire were presented, as given above, and no additional data were reported to support the use of such a formula.

An investigation was made by Sorokin (**53**) of the biographies of 1,600 leaders of labor and radical movements. Leaders were classified according to their socio-economic background and similar factors. The findings indicated that the majority of the labor and radical leaders come from families in which the paternal occupation is professional, business, or managerial.

A somewhat different survey approach was employed by Likert (**34**) in a study in which the relationship of managerial attitudes to the morale of life-insurance salesmen was investigated. In this study, approximately 300 insurance agents in twenty different agencies were interviewed systematically and each agent filled in a questionnaire con-

cerning his attitudes toward the firm, the management, etc. Each agent received a morale score based on the interview and one derived from the questionnaire, with the two scores correlating .85.

The agents indicated that the manager was the chief influence on their morale. The major characteristics of the manager which were considered important by the agents in this respect were: his attitude towards them; his personality in general; and his professional skill.

A number of investigators have attempted to develop test batteries to select executives or administrators. Among these may be listed Cleeton and Mason (12) whose battery of tests carries the rather elegant title of "Vocational Aptitude Examination for Sales, Technical, and Executive Groups." It was reported that this battery had been administered to personnel in executive positions, but no data were presented in this regard.

Uhrbrock and Richardson (58) administered a battery of nine tests to 163 supervisors. Ratings by superior executives were employed as criterion scores. Out of a total of 820 items in the nine tests, only 85 were found to have significant predictive value. The majority of the valid items were drawn from company information tests. The following personal history items were also significant in this study: age, schooling, and military service record.

In another industrial investigation Bridgman (8) performed a follow-up study of the success of 1,310 college graduates in the Bell Telephone System. The criterion involved was salary achieved in the company. The findings indicated that high scholarship, campus achievement, early graduation, and immediate employment in the Bell System were all significantly favorable factors for success in this company. Scholarship level during the college career appeared to be the most significant factor.

A study of 100 presidents and vice-presidents of "successful" companies has been reported by O'Connor (44). It was stated that five characteristics of executives were isolated and measured in the course of this research. The five items listed were: (1) large English vocabulary, (2) many aptitudes, (3) objective or extremely objective personality, (4) accounting aptitude, and (5) aptitude for first position. Descriptions of the instruments, conditions of administration and scoring, and data supporting this statement were not presented.

Beckman and Levine (4) studied the Allport A-S Reaction Test, the short form of a Personality Inventory (C-2), and a Directions Test in an attempt to predict supervisory ability. Various efficiency ratings were employed as criteria for an experimental and control group of approximately thirty cases each. On the basis of a correlation of .33 be-

tween supervisory ratings and scores on the Allport Test in this small sample, it was concluded that this test is of value in selecting executives.

Along these same lines, a study has been performed concerned with the administration of a mental alertness test to twenty-eight minor executives in a clothing manufacturing establishment (**31**). These men were rated by six superior executives, and it was reported that close agreement was found between the test scores and the ratings given the men by their superiors. As in the reports of many of these investigations the details necessary for critical evaluation were omitted.

In contrast to these studies reporting appreciable correlations of predictive instruments with ratings by supervisors for executives and administrators in industry are results reported, among a number of others, by Bingham and Davis (**7**). In this study, an intelligence test was administered to 102 business executives, and experience records containing personal information about the individuals were obtained from 73 of these men and employed as the criterion. No agreement between the two sets of data was found according to the authors.

Thurstone recently studied administrative ability in governmental work (**57**). A group of perceptual tests which had been shown to differentiate a small group of campus leaders from non-leaders was employed along with several new tests that were assumed to predict administrative ability. In the first phase of this study, a group of interns in governmental administration served as subjects. The ten interns with the highest supervisors' ratings of professional promise and success and the ten with the lowest ratings were selected and their test scores compared. The tests which were found to be most discriminating were Gottschaldt Figures, Street Completion, Kohs Block Designs (negatively), and a Two-Hand Tapping Test.

In another phase of the study, test scores were compared with a salary criterion corrected for age. For 127 administrators whose salaries ranged from less than $3,000 to $10,000 per year, the best single test was the linguistic score on the Psychological Examination of the American Council on Education. Gottschaldt scores yielded a prediction which was almost as accurate. A Classification Test involving categorization of cards with names printed on them, differentiated in that the higher paid administrators used fewer categories and fewer single card groups. On a test requiring that numerical estimates be made on the basis of common knowledge, such as the approximate population of the United States, the more successful men on the criterion did better, particularly in terms of the proportion of acceptable answers. The following scores on the Allport-Vernon Study of Values differentiated the two

groups: social, theoretical, economic, and religious. The successful men excelled on the first two and were inferior on the last two. They also had more masculine scores on the Terman-Miles Schedule for Masculinity-Femininity. On the Thurstone Vocational Interest Schedule, the administrators who had higher incomes made lower scores in physical science, physical activity, and commercial interests.

Richardson and Hanawalt (**50**) compared the Bernreuter Personality Inventory scores of 258 business men divided into groups of office holders and supervisors on the one hand and non-office holders and non-supervisors on the other. Office-holders and supervisors tended to be less neurotic, less introverted, more dominant, more self-confident, and more self-sufficient than the control groups and the norms. The differences were statistically significant for a majority of the measures.

The use of another procedure that purports to pertain to the selection of executives is illustrated in a German study by Luithlen (**35**). This investigation employed a test in which single words were printed on individual cards, there being two sets of cards one blue-bordered and the other red-bordered. Subjects were paired and one member of each pair was given the blue-bordered cards and the other the red-bordered ones. Each pair was requested to construct original sentences with the cards. In addition to general observations, the measures included time, number of constructed sentences, and color and position of the cards. On the basis of initial results, individuals were labelled as leaders or followers, and pairs of leaders and followers were given the same problem. It was found that when a leader and a follower were paired, the leader did most of the work, and when two followers performed, it took longer to reach the result, but there was cooperation. When the individuals were both leaders, fewer sentences were constructed. Statistical evaluation of the findings was not presented.

Studies of Scientific and Professional Personnel

Investigations similar to those listed above have been conducted concerning the characteristics of outstanding scientific and professional personnel. Several studies (**11, 47, 51, 60, 64**) have dealt with the characteristics of American inventors and have employed a questionnaire or some sort of biographical approach which yielded various normative data concerning the characteristics of these persons.

Typical studies of the characteristics of geniuses, utilizing similar methods, are contained in reports by Ellis (**16**) and Giese (**25**). Information from these investigations has indicated superiority over the general population for the specialized groups with regard to such items

as socio-economic background and education. They also agree in giving ages 25–35 as those of the first achievement for scientists, inventors, and literary men.

An interesting study has been performed by Schneider (**52**). The hypothesis under investigation in this study was that any group of distinguished men formed a "birth galaxy" associated with the cultural situation which, in turn, is related to a definite period of time affecting this entire birth group. The individuals studied consisted of approximately 200 English botanists born between 1700 and 1920. On the basis of published biographical information concerning these individuals, two outstanding periods of activity were uncovered. The conclusion drawn from the available data was that there occurred a period in the history of British botany when there were fewer persons born who became famous botanists than for the immediately preceding or following period. It appeared improbable that any racial factors were involved in the decline in the number of botanists from 1700 to 1800 since this period was marked by a general rise in the number of births of eminent Englishmen. Schneider also found that the fame of military leaders was associated with the number of conflicts in which the particular country was engaged.

A study has been reported by Flanagan (**19**), which pertains indirectly to leadership, concerning the validity of the 1940 edition of the Cooperative Test Service National Teacher Examinations. These examinations consisted of the following eleven sections: Reasoning, English Comprehension, English Expression, Current Social Problems, History and Social Studies, Literature, Science, Fine Arts, Mathematics, Professional Information, and Contemporary Affairs.

In order to validate these examinations 49 teachers in 22 school systems who had taken the tests were selected in such a manner as to maximize the range of scores (**20**). School supervisors then rated the teachers on a graphic rating scale consisting of 50 items, and students filled in an educational-report form concerning them. The product-moment correlation between the supervisors' ratings on the characteristic of "overall judgment of teachers' general effectiveness and desirability" and score on the examinations was .51. Positive correlations were found between all of the 50 rating items and total score on the tests. The lowest correlations were for the following ratings: teacher's health, physical appearance, and poise; energy, enthusiasm, and drive in school work; quality of speech and voice; sense of humor; congeniality of teacher's adjustment to associates; neatness of teacher's work in classroom; and integrity of teacher's character. It was found in this

small sample that the Current Social Problems section of the test yielded the highest correlation with the supervisor's over-all ratings.

A correlation significant at the five per cent level was found between total score on the examinations and the proportion of the students reporting a particular teacher's name in response to the question "Which teachers seem to have a broad knowledge of other subjects besides the one you had with them?" In addition, the correlation was significant at the five per cent level between a given teacher's score on the English Expression Test and the proportion of the same teacher's students who mentioned his name in response to the question, "Which teachers were most clear in presenting their ideas?"

Pre-School and Extra-School Studies Involving Children

The three major techniques which have been employed in the studies of leadership among children are: (1) observation, (2) nomination of companions for leadership positions, and (3) special test techniques and situations for identifying leaders. These three techniques are illustrated by several studies in the following paragraphs.

In a study by Parten (**45**) nursery school children were observed for 60 one-minute periods and their behavior recorded. There appeared to be a greater difference in leadership ability among the children than might be attributed to age alone according to the author. In addition, it was concluded that the leaders were above average in intelligence.

The nominating procedure was employed in a study at the Detroit Teachers College (**64**). In this investigation, more than 5,000 grammar-school children were asked to write the name of another child that they considered their best friend or would like to have as their best friend and the reasons for such a choice. The children were also asked to nominate the individuals they would like to have as leaders and present reasons. There were some outstanding differences reported between the reasons for selecting leaders as contrasted with those for selecting friends. Ability and achievement appeared to be the main reasons given for selecting a particular individual for a leader, and social characteristics, such as good sportsmanship, were much less significant for the leadership choices than they were for the friendship selections. It was found that a small number of children, about two to five in a given grade, received most of the votes.

In another study involving the nominating technique, Partridge (**46**) studied leadership in a sample of 226 Boy Scouts. The five-man-to-man rating technique was employed in which each subject rated every other individual in the group on leadership ability in groups of five cases with

the names presented in random order. The characteristics of those selected as leaders were recorded and compared with the ratings. It was found that leaders were generally superior individuals in such factors as height, weight, intelligence, age, and so forth. Wide individual differences appeared in the characteristics of leaders in different groups.

Partridge also studied the influence of leader attitudes upon group attitudes by having the group fill out a questionnaire and selecting items on which differences of opinion existed. A group discussion was held with a particular leader in charge and the group was then asked to vote again on the questions. The number who changed their opinions after learning the attitude of the leader was counted. The results of this study showed that the amount of influence leaders had upon group attitudes was greater than that of individuals who were not leaders. The differences were statistically significant

An early study by Terman (56) illustrates the use of special techniques in studying the problem of leadership. In this investigation groups of four children, boys and girls separately, were presented with ten pictures and objects on a large card and told that the experiment was a memory test and that they could look at the card for ten seconds and were then to answer questions concerning it. The answers were recorded in the order given. In a second series of tests, different subjects were used and the original groups were reformed. Each new group contained at least one individual who had been a leader in responding in the previous series and one who definitely had not been a leader. An analysis was made of the number of first, second, and third responses that were made to the stimulus cards in comparison with information concerning the children. The results on the small sample involved indicated that the leaders tended to be larger, brighter in school work, better looking, better dressed, more widely read, less emotional, more fluent, of more prominent parentage, and more daring.

Typical of a series of researches by Lewin and his associates is a study by Bavelas (3) concerning morale and the training of child leaders. The study dealt with the retraining of leaders with low morale. Three mediocre leaders on a WPA project, as judged by ratings, were selected and equated to a control group of three cases on the basis of age, sex, length of time on WPA, length of time on present WPA project, rating of technical skill, rating of leadership ability and relevant life history factors. All six individuals were tested by observing and recording their actual behavior with children on the job prior to and after training, and by recording the behavior of the children supervised by them. The experimental group was trained for two hours a day for a period of

twelve days by indicating to them in detail the attitudes, objectives, and philosophy of recreational group work. During this time the experimental and control groups continued their work with the children. During the fourth week both the trained and the non-trained individuals were tested again on the job by the same methods of observation and recording. Prior to training, leaders employed an authoritarian form of giving commands approximately 60% of the time, and the remaining portion of the time responded to the approach of the children. Following training, the experimental group used praise, invested the children with responsibility, and revealed their preferences rather than employing their previous approach. The number of children working together more than doubled for the experimental group while increasing approximately 40% for the control group.

Studies in the School Situation

As in several of the previous areas treated in this report, three general techniques have been employed in studies of leadership behavior in the school situation. These are the nominating procedure, the survey procedure, and the use of special test techniques. Several examples of each of these procedures are given below.

In a study which illustrates a type of nominating approach to the problem of leadership, Nutting (**43**) had a group of girls list the reasons for selecting certain individuals as captains for their gymnasium teams. An examination of the characteristics of these captains indicated that they were slightly above average in age, physical ability, and intelligence, and slightly below average in scholarship. They were noticeably high in "popularity."

The survey procedure is exemplified in a number of studies. Caldwell and Wellman (**10**) found that leaders in different activities exhibited different characteristics; for example, boys chosen as class presidents or other representatives were taller than average whereas the girl leaders were approximately average in height. Practically all leaders were high in scholarship, and although physical achievement was an important characteristic of athletic leaders, it did not appear to be relevant for other types of leadership. According to the results of a rating scale for extroversion, most of the leaders were more extrovert than introvert. Studies by Bellingrath (**5**) and Brown (**9**) employed similar procedures and yielded comparable results.

Zeleny (**62**) tried out several different procedures for selecting leaders in a group discussion situation. The first technique tried out was "identification by voice and appearance" which was reported to be suit-

able for speedy preliminary selection of leaders as judged by the supervisor of the group discussion. After groups of four or five individuals were formed, the "status-ranking technique" was used in which each individual ranked all others in his group and the average rankings were checked against special ratings of leadership by the group and by the group supervisor. This technique was reported to show some relationship between those selected by the group and the professor's ratings of the individuals on leadership qualities. The third procedure employed was the five-man-to-man rating technique. In Zeleny's situation a coefficient of .72 was reported between such ratings and the ratings by the faculty leader of the group discussion for a sample of twenty-one cases. The degree to which the faculty representative was familiar with the ratings of the students was not reported The final technique was known as "sociometric identification" in which each individual listed five choices for leader of the group discussion in order of preference. According to the author, by this technique one could determine those most chosen and also those with whom both leaders and followers would most like to work.

In another study by Zeleny (63), the five-man-to-man and status-ranking techniques were found to yield ambiguous validation data.

The extent to which several hundred girls desired to live and work with one another was rated by each member of the group in an investigation by Jennings (29). This procedure yielded a test-retest correlation after eight months for positive expression of choice of .65 and for rejection of .66. It appeared that those selected as leaders, to a very much greater extent than the average member of the group, constructively contributed to enlarging the social field for participation of other persons. The findings indicated the existence of many individual differences and an overlapping of characteristics between the leaders and non-leaders.

A factor-analysis study of the personality of high school leaders has been performed by Flemming (21). The criterion of leadership involved the assigning of points based upon positions of leadership held by 71 girls in grammar school. The predictors consisted of a check list of 46 traits to be filled out by the teacher for each girl, a ten-point scale concerning "the intensity of pleasant feeling" that each girl "subjectively associated with every other girl in her class," and a rating by the teacher on a ten-point scale of "the amount of personality each girl possessed." Thurstone's simplified method was employed for analyzing the matrix of intercorrelations between these two sets of data.

A correlation of .50 was found between leadership as defined above

and personality as rated by the teachers, and a coefficient of .33 between leadership and pleasingness of personality as rated by the girls.

Flemming concludes that there seemed to be four types of leadership ability in this group. Scores on the predictors for these selected traits yielded a correlation with the leadership criterion of .57 on the same sample. No cross application of the weights to a new sample was reported.

Thurstone (**57**) has performed a study in which the scores of 18 campus leaders on a battery of tests were compared with scores for a group of several hundred college non-leaders. One of the most consistent findings was that the campus leaders were less subject to visual illusions as determined by various tests. They gave indication of color dominance and superiority on a color-form sorting test, were less subject to brightness contrast, and showed a slower rate of alternation of ambiguous perspective in the Necker Cube. On the Rorschach, the leaders excelled in the total number of responses, were superior in the perceptual organization score, showed greater latency for color cards, and a smaller number of responses to the color cards in comparison to the black-and-white cards.

A comparison of the performance of college leaders in extra-curricular activities with the norms for the Bernreuter Personality Inventory showed significant differences favoring the former in dominance and the latter in introversion (**49**). Other findings were unreliable or inconsistent. These same authors have reported an item analysis of this inventory for the responses of 36 women college leaders and 45 non-leaders (**28**).

Military Leadership

A considerable amount of literature has accumulated during recent years concerned with non-empirical definitions of military leadership and speculations concerning the choice of leaders. More recently, military psychologists have devoted research efforts to certain phases of the problem of leadership. Typical information in this area is summarized below.

General. A series of books and pamphlets have been published by military writers concerning the problems of military leadership (**1, 2, 13, 37, 38, 39, 48, 67, 70**). Unfortunately, none of these tomes is based on empirically determined evidence and all of them reflect the personal opinions and speculations of the authors. Typical of this group of publications is a book by Ageton (**1**) concerning naval leadership. The author lists and discusses the following characteristics of leadership

which he considers to be of significance: simplicity, self-control, tact, honor, adherence to duty, and loyalty. The following types of catchwords and phrases are used to illustrate what the author believes are the characteristics of good leaders: practice what you preach, be cheerful, be a seaman, know your stuff, avoid careless criticism, etc.

A manual recently issued by the U. S. Bureau of Naval Personnel (67) presents a series of principles for younger officers to be used in selecting men for promotions and leadership positions. Certain of the principles appear to be based on widely accepted psychological postulates, but no evidence to support their use in this context is reported.

The main item of value that can be derived from the speculative reports on leadership appears to be an interpretation of the various military regulations which pertain to the problems of leading in military situations. These interpretations, discussions, and speculations concerning problems of leadership and how to behave as a military leader may be of value in helping to demark an area of research in which problems of leadership may be attacked.

An obvious starting place for a treatment of military leadership is a discussion of the procedures employed for selecting officers and officer candidates in various countries. Several of these are described in the following paragraphs.

Selection of officer candidates in the U. S. Army. Typical of the procedures employed in the armed services of the United States during World War II were those used in the U. S. Army. In this branch of the service, there were two types of commissioning: (1) direct commissioning from civilian life and (2) receiving a commission following attendance at an Officer Candidate School. Procedures for selecting U. S. aircrew officer candidates have been described in detail elsewhere and will be omitted here (for example, see the series of articles issued by the AAF Aviation Psychology Program, this Journal, 1943–1945).

In the early days of the war, a number of men were commissioned directly from civilian life into the U. S. Army. In practically all cases, these commissions were for specialists duties, particularly general administration. One factor that played a role in such commissioning was previous military experience, either Reserve or Regular Army. For such commissioning, there was an age minimum which was 30 in the early days of the war, and, in addition, the individual was required to pass a physical examination. In most cases the men were required to be college graduates. Other than these requirements and that of being qualified for some type of specialized duty there were practically no requirements. A Selection and Review Board examined the available in-

formation and accepted or rejected candidates. Occasionally men were interviewed by this Board. Practically all of the individuals commissioned in this manner did not participate in combat operations. There was no systematic follow-up study of the on-the-job performance of men so commissioned.

The other type of commission received in the United States Army consisted of attending an Officer Candidate School and, upon graduation, receiving a commission as a Second Lieutenant. The necessary prerequisites for such commissioning were as follows. First, it was necessary for the individual to make application and to meet certain requirements. Among these requirements were the meeting of a cut-off score on a general intelligence test, known as the Army General Classification Test, and the passing of a physical examination. Another necessary prerequisite for attending Officer Candiate School was that the individual receive the recommendation and approval of his immediate commanding officer who presumably was familiar with the work done by him. In addition, each man making application for attendance at Officer Candidate School had undergone a certain amount of basic training in an enlisted status. The final step in the sequence of application for admission to OCS was the meeting of an Officer Interview Board who interviewed each man individually and reviewed the various available information on his qualifications and finally approved or rejected his application. Many of the criteria employed by these boards were subjective in nature. If approved, the individual attended an Officer Candidate School and took various courses, differing with the branch of the service. If graduated, he received a commission as a Second Lieutenant, and, if eliminated, he was returned to duty as an enlisted man.

Selection of officer candidates in the British Army. An early report on officer selection procedures in the British Army is available in *Aviation Psychology Abstract* No. 56 (72). More recently a report has been released describing testing procedures for officer candidates in Great Britain at the close of the war (22). The basic element of the selection system was the War Office Selection Board (W.O.S.B.) which consisted of a President (Colonel), a Military Testing Officer (M.T.O.) (Major or Captain), and a commissioned psychologist and psychiatrist. Candidates were organized into groups of approximately eight for testing which lasted three days. The tests were of three kinds: pencil and paper, practical or field, and interviews. The printed tests included an educational and occupational questionnaire and a personal and medical one; three standard, general intelligence tests including mathematics and

figure analogy items requiring about 20 minutes each; and certain personality tests.

There were three practical tests: individual situations, command group situations, and leaderless group situations. In the case of the first the candidate might be asked to prepare and deliver a short talk to fellow candidates on group morale. In the second type of test the candidates took turns organizing their groups to solve an immediate, practical problem. The last situation was a test in which the group was faced with a task such as escaping across an electrically charged fence without the M.T.O.'s having appointed a leader.

The procedures for scoring these various practical tests were not described, but were presumably subjective in nature. The difficulties in standardizing and objectifying such field measures in order to achieve adequate reliability and validity are obvious although the attempt to use such techniques is noteworthy.

A series of interviews were also employed including a psychiatric interview, an "officer quality" interview carried out by the Board President, and a technical interview for qualification in special service branches performed by a specialist in the particular field. A final Board Conference was held to meet the candidate and discuss his potentialities as officer material.

Data comparing ratings of the candidates during training with test scores were reported, indicating superiority of the men selected by these procedures over those who had entered service prior to the initiation of these selection techniques. Without additional information concerning the circumstances of the validation, particularly about the criterion of proficiency, it is not possible to evaluate the findings.

Selection of officer candidates in the German Army. A detailed description of German testing procedures for officer candidates as of 1941 is presented in a volume issued by the Committee for National Morale (17). The selection techniques are reported to include a life history examination, expression analysis (facial, body, voice, appearance, and handwriting), mental capacity as measured by intelligence and interests tests, including projection and completion examinations, and action analysis. The latter consisted of two tasks, the first being the "command series" in which the candidate was given a series of orders to execute; his behavior was observed and recorded by the testers. The second phase of action analysis was the leadership test in which a group of infantry soldiers were placed under the command of the candidate who supervised the execution of a pre-assigned task, such as the assembling of a prefabricated bridge. The behavior of the candidate in executing

this task was noted, and the men working for him were questioned. The similarity of these methods to those of the British is apparent. In addition, an apparatus test was employed involving differential reaction to a number of levers in response to different patterns of red, white, and blue lights and to lights with circles and squares which were successively lit. The candidate's errors and speed of response were automatically recorded, and his behavior while taking the test was observed. A test of choice reaction to auditory stimuli with the Rupp falling rod apparatus was also employed.

The leadership examination was conducted at an Armed Forces Testing Station by a board of examiners consisting of an Army Colonel, a medical officer, and three psychologists. Two full days were devoted to testing for the Army and two and one-half for the Air Forces. The general tests were given to groups of four and five at a time, and the interviews were conducted individually by the psychologists. Round-table discussions with the candidate were held and, during the one day interval interspersed between the two days of testing, the candidate was observed. Procedures for scoring the tests, for combining the scores statistically, and for follow-up studies of predictive efficiency apparently were not reported by the German psychologists.

Fitts (18) has recently reported the details of German selection procedures, particularly those in the Air Force, based on interviews with the military psychologists involved. Several items pertinent to the present discussion are contained in his report. For example, he points out how political, practical, and scientific factors contributed to the breakdown of the German Aviation Psychology Program. Of considerable significance in the present connection is the fact that Fitts was unable to locate any validation data which appeared to meet ordinary scientific criteria. The German program is a beautiful example of the uselessness of elaborate testing techniques hand in glove with complete disregard of their necessary concomitants—standardization, objectification, and validation.

Research proposals and studies. A number of proposals have been made and several researches undertaken dealing directly or indirectly with military leadership (see, for example, 6, 14, 15, 30, and 69.) Typical of these suggestions and investigations are the following items.

A group at Harvard (61) has proposed the use of a physical fitness test, the Harvard Step Test; an interview; and somatotype measures for selecting combat officers. No over-all comparison of combined scores with criteria of officer efficiency was presented.

Murray and Stein (42) have made suggestions concerning the selec-

tion of combat officers. Their emphasis and philosophy appear to be clinical in nature, not unlike the approach of the German psychologists reported previously. The measures included an interview, Murray's Thematic Apperception Test administered individually and to groups, a Construction Test similar to the British practical tests and the German leadership tests, and a complex perceptual-motor test taken under verbal stress. A group conference is held at the end of testing for a terminal evaluation based on test performance and clinical judgments of leadership characteristics. Validation data were not reported for the use of this test battery.

A description of a number of potential measures for selecting combat leaders has recently been presented (**36**). These include an interview stressing participation in outdoor activities and a background of self-confidence, several perceptual-motor tests, and a printed test of military adaptability consisting of a series of multiple choice items in the form of descriptions of situations based upon actual combat events.

The results of a number of military studies pertaining to leadership are not available for release at the present time, but several in the Army Air Forces have pointed out a number of deficiencies in the reliability and validity of the criteria involved (**66, 68, 71, 73**). The findings of these investigations indicate a lack of agreement between independent judges when rating on over-all performance, and recommendations were made in these studies for the use of ratings of work samples of performance in specific situations which have been directly observed as criteria. The difficulties of finding measures that correlate highly with success in Officer Candidate Schools have been reported by other investigators (**23, 24, 74**).

A few investigations of military leadership that have been released are treated in the following paragraphs. Several reports of studies have been issued by the Information and Education Division of Army Service Forces concerned with morale in relation to problems of leadership in the U. S. Army (**65, 69**).

In the first study privates in two regiments of an infantry division took an attitude questionnaire during their first few weeks in the Army, and the results of the study were reviewed to see how the morale attitudes of men who rated promotions compared with those of other men. The study showed that the privates who became non-commissioned officers some months later were likely to differ from their buddies in the following regards: better disciplined, more self-confident, more likely to feel that what they were doing in the army was worthwhile, and more favorable attitudes toward their commissioned and non-commissioned officers. The statistical significance of these findings was not reported nor were the items on which differences were low or negative presented.

The fact that men who get along with their officers tend to be promoted is not surprising. As another part of the study, line noncoms in the two regiments (assumed to be the best enlisted infantrymen) were compared with privates and pfc's on such factors as education, AGCT scores, physical characteristics and mechanical aptitude. The noncoms were generally found to have more education, intelligence, and mechanical aptitude, and to be slightly taller and heavier than the privates and pfc's. In all the above cases, however, the differences between noncoms and privates and pfc's were small. It was evident that the most striking differences between the two groups were in their morale attitude.

Another study (69) reports a check-list of company leadership practices. This report is based on a study of practices among 34 Army Service Forces companies in the continental United States. The problem was the relationship between company leadership practices and company morale. Twelve selected companies answered eighteen questions concerning leadership practices in their own outfits. Six of the companies were all rated high in morale by their post or battalion commander, their company officers, and their enlisted men. The six other companies were all rated low in morale by corresponding judges. The companies rated highest in morale by these judges were favorably rated by their own men on nearly all company practices. Those rated lowest fared very poorly in this respect. The items on which two-thirds or more of the men expressed favorable opinions, in all six of the companies rated highest in morale, involved an expression of interest in the men by the officers.

The Multiple-Choice Group Rorschach Test has been administered to 56 officers selected by unit commanders as being actually excellent officers (30). Their scores were compared with those of 257 officer candidate students enrolled in three classes. The scores for the excellent officers were approximately the same as those for the officer candidates. The difference was not statistically significant. Forty-five per cent of the excellent officers and 35% of the officer candidates had four "poor responses" which was set as the critical score for initial screening purposes by Harrower-Erickson. Results on a Health Inventory, a Psychasthenic Inventory, and a Level of Aspiration Inventory also revealed no significant differences between the two groups.

Several studies dealing with military leadership were performed by the Applied Psychology Panel of the National Defense Research Committee during the war (23, 24). In a preliminary study in Infantry and Field Artillery Officer Candidate Schools it was found that individual raters tended to be consistent from one week to another when members of each platoon rated one another. Correlations between the platoon members' ratings and platoon leaders' ratings were also high, but appeared to be spurious since the member ratings were available to the platoon leader before his ratings were reported.

In follow-up studies data were analyzed dealing with the forecasting of success or failure by various predictors in these same schools. It was concluded that none of the items obtainable before or upon entrance to Officer Candidate School is highly significant of later performance; quality of academic work is significantly related to success or failure in O.C.S.; and platoon leaders' ratings of leadership correlate highly with success as would be expected in this situation where a man is eliminated if his ratings by the platoon leader are consistently low.

In another study, 176 men on combat duty were rated by their superior officers on a five-point scale in a situation that indicated that the rating officers were taking account of the actual performance and were not rating simply on general impressions. There was some spread in the ratings with 13% of the 176 officers receiving superior ratings, 49% excellent, 23% very satisfactory, 10% satisfactory, and 5% unsatisfactory. These combat ratings were compared with final company ratings in O.C.S., with Army General Classification Test scores, and with age. The conclusions drawn from this study were as follows:

1. Combat efficiency is not very closely related to ratings of leadership obtained in O.C.S. The correlation between the two sets of ratings was .15 with a standard error of .075.
2. It is reasonably clear that above a certain desirable minimum, intelligence as measured by AGCT has little relevance to combat performance.
3. At the present time the Infantry School is eliminating a large number of those who would probably be unsuccessful officers. In addition, there appeared to be a practically zero relationship between age and combat ratings in this study.

One study that is of considerable interest in the present connection has recently been undertaken by the Personnel Research Section, Classification and Replacement Branch, Adjutant General's Office (6), involving the development of instruments for selecting officers for positions in the post-war Army. Certain of the instruments derived from this study are now being employed in selecting Regular Army personnel from available applicants. Five kinds of procedures were employed in this study.

The first of these was the Officer Evaluation Report (OER) which was developed as an improvement over the procedure of averaging previous efficiency ratings. The OER consists of five sections, the first two of which were designed to improve its objectivity. The third section contains five items which appear to measure all three of the major named rating factors—leadership, sense of duty, and stability which were derived on the basis of a factor analysis of War Department and AAF efficiency forms. The last two sections of the OER concern recommendations for the Regular Army and for the Officer Reserve Corps and an over-all rating giving the officer's standing among previously known officers of the same grade.

The second instrument which was developed is the New Interview Board.

Its main function is to evaluate the officer's ability to get along with people in line and staff functions. The interview form provides two work sheets upon which to record observations, reactions, and ratings on three areas called bearing or manner, voice and language, and personality traits. The manual and prepared work sheets attempt to emphasize the social nature of the interview and minimize the influence of prejudice by forbidding prior acquaintance of any one of the five-member board with the applicant, any prior knowledge of general experience or background, or any attempt to explore or evaluate any phases of these not brought out by discussion during the interview.

The third primary measure was a Biographical Information Blank requiring the reporting of background information of a personal, educational, and occupational nature in combination with self-description of traits, attitudes, and habits. The blank is based upon the previous experience of the National Research Council Committee on Selection and Training of Aircraft Pilots, National Defense Research Committe, the Worker Analysis Section of the Occupational Research Program of USES, and the U. S. Navy Department.

Two other instruments were also developed in the course of this project to check general learning aptitude and educational achievement: (1) an Officer Classification Test, a test of general intelligence, and (2) a General Survey Test, a measure of educational achievement.

Approximately 15,000 officers participated in the field trials in about fifty Army installations. The criterion involved a nomination procedure as follows. Approximately 15,000 officers were brought together in small groups of 15 to 30 who were well enough acquainted with each other's work to evaluate it. Each officer was requested to prepare two lists of the officers in his group, those High and those Low in general all-around value to the Army, listing them from highest to lowest, next highest, next lowest, etc. Each officer was requested to designate the five officers of the group most closely medium or middle with respect to over-all value. Only officers who fell clearly into one of the three groups—High, Middle, or Low—in the sense of virtually perfect agreement with their fellow officers (never more than one dissenting vote for high or low officers, and never more than two dissenting votes in the case of middle officers) were employed for further study. In addition any officer placed over one group away by the Commanding Officer was eliminated. A final sample of 3,000 cases was employed which appeared to be selected with 1,000 in the Top group, 1,000 in the Middle group, and 1,000 in the Low group.

The Biographical Information Blank and the Interview were combined with the Officer Evaluation Report by statistical weighting procedures to secure a Combined Point Index. Pearsonian correlation coefficients were computed by arranging the criterion variable in the three categories, H, M, and L, and each predictor as a continuous variable. These coefficients were as follows: Officer Evaluation Report .60, New Interview .39, Biographical Information Blank .35, Combined Point Index .67, Previous Efficiency Report .45, and Traditional

Army Board .09. It should be noted that these coefficients may have been increased by selecting samples of 1,000 for the Top, Middle, and Low groups from the total sample of 15,000. The degree to which the Commanding Officers' ratings were weighted in the Officer Evaluation Report was not stated, but it appears likely that this factor played an important role. Substantial agreement between ratings by the C.O. and by fellow officers was to be expected. Since the OER had the highest validity, and the other measures when combined with it increased its correlation with the criterion only .07, these questions suggest the necessity for a further examination of the nature of the criterion here employed.

Neither the Officer Classification Test nor the General Survey Test was related to the High, Middle, or Low criterion classification of the men, a finding which might have been expected since college graduation had been a prerequisite for most men directly commissioned and a score of 110 on the Army General Classification Test had been required for admission to O.C.S. or cadet training. The General Survey Test showed moderate agreement with educational level reached. A cut-off score on the Officer Classification Test was recommended as a first hurdle before the Combined Point Index is applied. With regard to the General Survey Test, it was stated that while no cutting score on this instrument was recommended, it was believed that this test might be a better requirement than a general educational prerequisite.

The general conclusion drawn from this investigation was that the Combined Point Index will tend to select officers satisfactorily on the basis of past and present performance.

A research project concerned with combat leadership has recently been reported from the AAF Aviation Psychology Program (**15**). While the findings of these several studies cannot be released at the present time, the techniques are worth mentioning briefly. The first technique involved the collection of anecdotes dealing with the types of behavior exhibited in combat by men designated as successful or unsuccessful leaders or by those who assumed this role in cases of emergency. These materials were systematized and categorized, and secondly a rating scale was derived which served as a check on the results of the previous method. The third technique involved the use of one of the more objective of the available criteria of leadership, namely, promotions. Data from a number of selection devices were compared with this criterion.

Discussion and Conclusions

As an over-all generalization it appears that Viteles' (**59**) statement in 1932 concerning research on the selection of executives still holds for the general area of leadership: the record of accomplishment is not a brilliant one. No single trait or group of characteristics has been iso-

lated which sets off the leader from the members of his group. Several writers (**32, 40, 41**) in summarizing the literature in specific areas have stressed the cultural and situational determination of leadership. They have also pointed out the existence of wide individual differences within a given group as well as between groups.

Advances in methodology in this field are definitely not striking. Three techniques have generally been employed: nomination of members of the group for positions of leadership, survey of the characteristics of outstanding individuals involving the use of questionnaires or published biographical information, and the use of selection test techniques for identifying leaders. Progress has not been made in the development of criteria of leadership behavior, nor in the setting-up of an adequate working definition of the concept to guide research in the isolating of leadership traits.

The situation does not appear to be a particularly happy one with regard to the deriving of general principles or of setting up a systematic theory of leadership from the available information. A few statements may be set forth, however, that appear to hold for the findings of a number of the investigations reviewed; this list should be thought of as a series of hypotheses for further investigation.

1. Leadership is specific to the particular situation under investigation. Who becomes the leader of a given group engaging in a particular activity and what the leadership characteristics are in the given case are a function of the specific situation including the measuring instruments employed. Related to this conclusion is the general finding of wide variations in the characteristics of individuals who become leaders in similar situations, and even greater divergence in leadership behavior in different situations.

2. In practically every study reviewed leaders showed some superiority over the members of their group in at least one of a wide variety of abilities. The only common factor appeared to be that leaders in a particular field need and tend to possess superior general or technical competence or knowledge in that area. General intelligence does not seem to be the answer for, as one writer (**32**) has pointed out, public leaders have ranged all the way from dull normal to genius.

3. Leaders tend to exhibit certain characteristics in common with the members of their group. Two of the more obvious of these characteristics are interests and social background.

4. Certain past history or background items appear to characterize leaders in certain activities. There is a widespread but vague hint that certain poorly defined personality traits characterize individuals holding positions of responsibility. It is practically impossible to evaluate these suggestions without additional research.

5. A number of studies suggest superiority of leaders over those in their group in physique, age, education, and socio-economic background, but the need for further research in this connection is evident.

Summary

This article reviews the findings and techniques of a number of investigations of leadership in industry and government, the professions, school, and military situations. Special emphasis is given to the procedures for selecting officer personnel in the armies of the United States, Great Britain, and Germany; research studies of military personnel are stressed. Findings that appear to be common to a number of investigations are presented as hypotheses or possible principles of leadership behavior.

BIBLIOGRAPHY

1. Ageton, A. A. *Naval leadership and the American bluejacket.* New York: McGraw-Hill, 1944.
2. Andrews, L. *Military manpower.* New York: Dutton, 1920.
3. Bavelas, A. Morale and the training of leaders. In G. Watson (Ed.), *Civilian morale.* Boston: Houghton Mifflin, 1942.
4. Beckman, R. O., & Levine, M. Selecting executives, an evaluation of three tests. *Person. J.*, 1930, **8**, 415–420.
5. Bellingrath, C. C. *Qualities associated with leadership in extra-curricular activities of the high school.* Teachers College, Columbia Univ., Contr. Educ., No. 399. New York: Teachers College, Columbia Univ., 1930.
6. Bellows, R. M. Evaluating officer performance. Mimeographed report presented at the joint Army-Navy-OSRD Conference, August 16, 1945.
7. Bingham, W. V., & Davis, W. T. Intelligence test scores and business success. *J. appl. Psychol.*, 1924, **8**, 1–22.
8. Bridgman, D. S. Success in college and business. *Person. J.*, 1930, **9**, 1–19.
9. Brown, M. *Leadership among high school pupils.* Teachers College, Columbia Univ., Contr. Educ., No. 559. New York: Teachers College, Columbia Univ., 1933.
10. Caldwell, O. W., & Wellman, B. Characteristics of school leaders. *J. educ. Res.*, 1926, **14**, 1–13.
11. Carr, L. J. A study of 137 typical inventors. *Publ. Amer. sociol. Soc.*, 1929, **23**, 209.
12. Cleeton, G. U., & Mason, C W. *Executive ability, its discovery and development.* Yellow Springs, Ohio: Antioch Press, 1934.
13. Cope, H. F. *Command at sea.* New York: Norton, 1943.
14. Cowley, W. H. The traits of face-to-face leaders. *J. abnorm. soc. Psychol.*, 1931, **26**, 304–313.
15. Crannell, C. W., & Mollenkopf, W. G. Combat leadership. In *Psychological research on problems of redistribution*, AAF Aviation Psychology Program Report No. 14. Staff, Psychological Division, Hq. AAF Personnel Distribution Command, Louisville, Ky., February, 1946.
16. Ellis, H. *A study of British genius.* Boston: Houghton Mifflin, 1926.
17. Farago, L. (Ed.) *German psychological warfare.* New York: Committee for National Morale, 1941.
18. Fitts, P. M. German applied psychology during World War II. *Amer. Psychologist*, 1946, **1**, 151–161.

19. FLANAGAN, J. C. A preliminary study of the validity of the 1940 edition of the National Teacher Examinations. *Sch. & Soc.*, 1941, **54**, 59–64.
20. FLANAGAN, J. C. An analysis of the results from the first annual edition of the National Teacher Examinations. *J. exp. Educ.*, 1941, **9**, 237–250.
21. FLEMMING, E. G. A factor analysis of the personality of high school leaders. *J. appl. Psychol.*, 1935, **19**, 596–605.
22. GARFORTH, F. I. DE LA P. War Office selection boards (O.C.T.U.). *Occupational Psychol.* (London), 1945, **19**, 96–108.
23. GARRETT, H. E., & LIGON, E. M. Study of combat leadership. Mimeographed memorandum, Applied Psychology Panel, National Defense Research Committee, March, 1944.
24. GARRETT, H. E. Second report on combat leadership. Mimeographed memorandum, Applied Psychology Panel, National Defense Research Committee, June, 1944.
25. GIESE, F. The public personality. A statistical study of the intellectual leaders of the day. *Beih. Z. angew. Psychol.*, 1928, No. 44.
26. GOWIN, E. B. *The executive and his control of men.* New York: Macmillan, 1915.
27. GOWIN, E. B. *Selection and training of the business executive.* New York: Macmillan, 1918.
28. HANAWALT, N. G., RICHARDSON, H. M., & HAMILTON, R. J. Leadership as related to Bernreuter personality measures: II. An item analysis of responses of college leaders and non-leaders. *J. soc. Psychol.*, 1943, **17**, 251–267
29. JENNINGS, H. H. *Leadership and isolation.* New York: Longmans, 1943.
0. JENSEN, M. B., & ROTTER, J. B. The validity of the Multiple Choice Rorschach Test in officer candidate selection. *Psychol. Bull.*, 1945, **42**, 182–185.
31. KORNHAUSER, A. W., & KINGSBURY, F. A. *Psychological tests in business.*, Chicago: Univ. Chicago Press, 1924.
32. KROUT, M. H. *Introduction to social psychology.* New York: Harper, 1942.
33. LAIRD, D. A. *How to use psychology in business.* New York: McGraw-Hill, 1936.
34. LIKERT, R. *Morale and agency management.* Hartford, Conn.: Life Insurance Sales Research Bureau, 1940.
35. LUITHLEN, W. F. The psychology of initiative and leadership traits. *Z. angew. Psychol.*, 1931, **39**.
36. MEIER, N. C. *Military psychology.* New York: Harper, 1943.
37. MILLER, A. H. *Leadership.* New York: Putnam, 1920.
38. MUNSON, E. L., JR. *The management of men.* New York: Holt, 1921.
39. MUNSON, E. L., JR. *Leadership for American army leaders.* Washington: Infantry J. Publ., 1942.
40. MURPHY, G., & MURPHY, L. B. *Experimental social psychology.* New York: Harper, 1931.
41. MURPHY, G., MURPHY, L. B., & NEWCOMB, T. M. *Experimental social psychology.* (Rev. Ed.) New York: Harper, 1937.
42. MURRAY, H. A., & STEIN, M. Note on the selection of combat officers. *Psychosom. Med.*, 1943, **5**, 386–391.
43. NUTTING, L. R. Some characteristics of leadership. *Sch. & Soc.*, 1923, **18**, 387–390.
44. O'CONNOR, JOHNSON. *Psychometrics.* Cambridge: Harvard Univ. Press, 1934.
45. PARTEN, M. B. Leadership among preschool children. *J. abnorm. soc. Psychol.*, 1933, **27**, 430–440.

46. PARTRIDGE, E. D. *Leadership among adolescent boys.* Teachers College, Columbia Univ. Contr. Educ., No. 608. New York: Teachers College, Columbia Univ., 1934.
47. RASKIN, E. Comparison of scientific and literary ability: a biological study of eminent scientists and men of letters of the nineteenth century. *J. abnorm. soc. Psychol.* 1936, **31**, 20–35.
48. REED, P. B., JR. *Personal leadership for combat officers.* New York: McGraw-Hill, 1943.
49. RICHARDSON, H. M., & HANAWALT, N. G. Leadership as related to the Bernreuter personality measures: I. College leadership in extracurricular activities. *J. soc. Psychol.*, 1943, **17**, 237–249.
50. RICHARDSON, H. M., & HANAWALT, N. G. Leadership as related to the Bernreuter personality measures: III. Leadership among adult men in vocational and social activities. *J. appl. Psychol.*, 1944, **28**, 308–317.
51. ROSSMAN, J. Study of childhood-education and age of 701 inventors. *J. Pat. Off. Soc.*, 1935, **17**, 411–421.
52. SCHNEIDER, J. The cultural situation as a condition for the achievement of fame. *Amer. sociol. Rev.*, 1937, **2**, 480–491.
53. SOROKIN, P. A. Leaders of labor and radical movements in the United States and foreign countries. *Amer. J. Sociol.*, 1927, **33**, 382–411.
54. STARCH, D. *How to develop your executive ability.* New York: Harper, 1943.
55. TAUSSIG, F. W., & JOSLYN, C. S. *American business leaders.* New York: Macmillan, 1932.
56. TERMAN, L. M. A preliminary study in the psychology and pedagogy of leadership. *Pedag. Sem.*, 1904, **2**, 413–451.
57. THURSTONE, L. L. *A factorial study of perception.* Chicago: Univ. of Chicago Press, 1944.
58. UHRBROCK, R. S., & RICHARDSON, M. W. Item analysis: the basis for constructing a test for forecasting supervisory ability. *Person. J.*, 1933, **12**, 141–154.
59. VITELES, M. S. *Industrial psychology.* New York: Norton, 1932.
60. WINSTON, S. Bio-social characteristics of American inventors. *Amer. sociol. Rev.*, 1937, **2**, 837–849.
61. WOODS, W. L., BROUHA, L., & SELTZER, C. C. *Selection of officer candidates.* Cambridge: Harvard Univ. Press, 1943.
62. ZELENY, L. D. Characteristics of group leaders. *Sociol. soc. Res.*, 1939, **24**, 140–149.
63. ZELENY, L. D. Objective selection of group leaders. *Sociol. soc. Res.*, 1939, **24**, 326–336.
64. *How children choose friends.* [Anon.] Detroit: Society for the Scientific Study of Character, Detroit Teachers College, 1929.
65. *A check-list of leadership practices.* Research Branch, Information and Education Division, Army Service Forces. In *What the Soldier Thinks*, April 1945, No. 5, pp. 38–39.
66. *Interrelationships of the scores making up the flight officer composite.* Psychological Research Unit No. 3 Santa Ana, Calif. *Research Bulletin S44-9*, March, 1944 (Restricted).
67. *Manual for practical development of leadership qualities.* United States Navy Department, Bureau of Personnel. Washington: Government Printing Office, Oct., 1944.
68. *The measurement of officer quality in the AAF officer candidate school.* Psychological Research Unit, San Antonio, Texas. *Research Bulletin, T44-19*, Oct., 1944.
69. *Morale attitudes of superior infantrymen in training.* Research Branch,

Information and Education Division, Army Service Forces. In *What the Soldier Thinks*, April, 1944, No. 5, pp. 12–13.

70. *Naval leadership.* (4th Ed.) Annapolis: U. S. Naval Institute, 1939.
71. *The relationship of various tests to officer quality criteria available at the AAF Adm. OCS, Miami Beach Florida.* Psychological Section, Office of the Surgeon, HQ AAF Training Command. *Research Bulletin Hq. 44-29*, June, 1944 (Restricted.)
72. *Selection of officer candidates in Great Britain.* Psychological Branch, Office of the Air Surgeon, HQ AAF. *Aviation Abstract Series*, December, 1942, No. 56 (Restricted).
73. *Study of officer validity scores used in connection with the Flight Officer Act.* Psychological Research Unit, San Antonio, Texas. *Research Bulletin T44-14*, August, 1944 (Restricted).
74. *A study of prediction of success in Marine officer candidate school by ratings.* Psychological Branch, Office of the Air Surgeon, HQ AAF. *Aviation Psychology Abstract Series*, March, 1945, No. 157 (Restricted)

COMMENT ON "THE VALIDITY OF PERSONALITY QUESTIONNAIRES"

CHRISTIAN PAUL HEINLEIN
Florida State College for Women

In the september 1946 issue of the *Psychological Bulletin*,* Dr. Albert Ellis reviews a series of studies involving the use of personality inventories of different kinds. After considering the arguments for and against the various practices of validating personality questionnaires, he proceeds to evaluate the validity of selected research results by means of the following verbal categories assigned to arbitrary segments of the familiar Pearsonian scale of correlation coefficients:

1. *Negative* validity (*r*'s of from .00 to .19)
2. *Mainly negative* validity (*r*'s of from .20 to .39)
3. *Questionably positive* validity (*r*'s of from .40 to .69)
4. *Mainly positive* validity (*r*'s of from .70 to .79)
5. *Positive* validity (*r*'s of from .80 to 1.00)

In arriving at this five-fold verbal classification of validity, Dr. Ellis remarks:

> "Since holding personality test evaluations to terms of *D* or *E* rather than *r* would probably be *unfair* (italics mine) at the present stage of their development, we shall, in this review, usually evaluate the reported coefficients of correlation in terms of the conventional estimations given them in the consideration of psychological and educational tests."

What, may I ask, could possibly be more "unfair" than to delude those readers who are unfamiliar with the mathematical derivation of *r* into believing that the phrases "mainly positive validity" and "positive validity" impart operational significance to the results of personality questionnaires? The five verbal categories which Dr. Ellis has coined are of the nature of semantic blabs, without operational referents. Arbitrary ranges of correlation coefficients, which determine the limits of application of the five verbal classifications, can hardly be regarded as operational referents. The simple exchange of a verbal sign for a numerical sign is not the discovery of an operational referent for the concept of validity.

In his discussion of correlational techniques utilized for the purpose

* Ellis, Albert. The validity of personality questionnaires. *Psychol. Bull.*, 1946, 43, 385–440.

of validating personality questionnaires, it is obvious that Dr. Ellis cognizes r as a trigonometric function; otherwise it is doubtful whether he would have mentioned the need of converting r into E; namely, $1-(1-r^2)^{1/2}$. To strengthen this view, we have his expressed conviction that "many correlations reported on personality-test validity experiments should not be shown in terms of r at all." If one were in a position to demonstrate that an r of a given size is an unequivocal index of functional dependence between two arrays of values normally distributed, and if, further, one were in a position to predicate empirically by virtue of unambiguous knowledge of the causative conditions which operate to provide a specified range of effects in response, then one would need the assurance of an r of at least .86 in order to bring about the reduction of the probable error of predication to one-half the probable error of guessing. In spite of this fact, Dr. Ellis selects an r of .80 as the lower limit of his highest category of validity. Conversion of the five successive verbal categories of validity into ranges of E provides the following highly questionable series:

1. *Negative* validity, E's from .00 to .01
2. *Mainly negative* validity, E's from .02 to .07
3. *Questionably positive* validity, E's from .08 to .27
4. *Mainly positive* validity, E's from .28 to .38
5. *Positive* validity, E's from .40 to 1.00

It can readily be seen from this type of conversion—a conversion which Dr. Ellis favors but, paradoxically enough, regards as "unfair"—that the fifth category covers an extension of value greater than that of the first four categories. If the above series seems questionable for the purpose of establishing categories of validity, then let it be remembered that the selection of the r ranges from which the E ranges have been derived is equally questionable.

Unfortunately in the history of psychometrics, the coefficients r and E as indices of numerical communality have been grossly misunderstood and widely abused. The method of Pearsonian rectilinear correlation, which ignores the function of a time-axis and utilizes fixed individual deviations in a static matrix of paired numerical values, has been falsely identified with the method of concomitant variation. In the latter method, a controlled, isolated unit event A is varied by an observable, measurable amount to ascertain its determining *effect* upon an observable, macroscopically constant unit event B. Only by a fantastic stretch of imagination can one attribute cause-and-effect relationship or a determining, predictive function to an r as such, irrespective of the size of the r. By a convenient act of extrapolation, one may choose to

ascribe a certain level of validity to an *r* of a given size. Such practice may be one of expediency or wishful thinking, but it certainly is not one of sound, scientific procedure. An adequate proof of a pragmatic kind is sorely needed to demonstrate the efficacy of *r* as an index of validity for a specific type of personality questionnaire or test situation. The arbitrary verbal classifications introduced by Dr. Ellis do not constitute this kind of required proof. At most, they merely add to the extensive confusion that exists in the literature on test-validation.

DISCUSSION OF HEINLEIN'S COMMENT ON "THE VALIDITY OF PERSONALITY QUESTIONNAIRES"

ALBERT ELLIS
New York City

Heinlein's comment on my review of *The Validity of Personality Questionnaires* contains several points with which I concur and several with which I must take issue. I shall consider it paragraphically.

In his first paragraph, Heinlein states that I proceed "to evaluate the validity of selected research results by means of . . . categories assigned to arbitrary segments of the familiar Pearsonian scale of correlation coefficients." There are at least three implications here that may be misleading:

1. That I utilized *selected* research results;
2. That I evaluated these results only in terms of correlation coefficients;
3. That the verbal categories I assigned in evaluation of correlation coefficients were quite *arbitrary*.

Actually, the facts are these:

1. I reported all research results that could be found in an extensive review of the literature.
2. I evaluated the correlations of only those studies reporting findings in terms of *r*. Many studies reported critical ratios or other clear-cut statistical measures which could be accepted without any special evaluations.
3. The verbal categories I applied to correlation coefficients were not arbitrarily chosen. They have been used as a rule of thumb in the psychological literature for many years (e.g., Garrett, 4, p. 342), are endorsed by many authorities, and have been widely applied in psychology classes.

A valid criticism which Heinlein fails to make of my verbal categories is my inadvertently misnaming the first two of them. Strictly speaking, there is no such thing as "negative validity." What I call "negative validity" in my article should really be termed "negligible validity," and what I call "mainly negative validity," should be termed "low validity."

In the second paragraph of his comment, Heinlein quotes me to the effect that holding personality test evaluation to terms of *E* rather than *r* would be unfair at the present stage of their development. Unfortunately, he italicizes *unfair* when the correct emphasis should be on *the present stage of their development.* Throughout his comment he falsely emphasizes *unfair*—henceforth wholly taken from its context—to make it appear as if I am contradicting myself. For it certainly would be a

contradiction were I to say that using E is an *unfair* statistical device—and then also to say that I favor its use. What my phrasing, correctly emphasized, means of course is that the use of E is perfectly fair, statistically and strictly speaking, but that because personality questionnaires are still (I trust) in their infancy or adolescence, taking r's obtained in validity experiments and reformulating them in terms of E would be unduly rigorous *at the present time.* Consequently—in order to give present-day personality inventories every possible benefit of the doubt—I consistently evaluated the reported r's for validity in a conventional manner—and, as Heinlein points out, in a generous manner indeed. Whether I should be that generous twenty-five or fifty years from now is highly dubious.

Heinlein's third paragraph is concerned with taking me to task for "deluding" my readers into believing that my verbal categories impart operational significance to the results of personality questionnaires. I naturally tried to do no such thing. First of all, I was not concerned with "the results of personality questionnaires," but with the results of experiments attempting to validate questionnaires—which is quite a different thing. Secondly, I was evaluating r's only *because they had been reported by investigators*, and not because I personally believe that validity experiments should be analyzed in terms of r. I heartily agree with Heinlein that "the simple exchange of a verbal sign for a numerical sign is not the discovery of an operational referent for the concept of validity." I must remind him, however, that since r *is* still being constantly used as a measure of test validity, and since evaluations of reported r's *are* bound to vary among different observers, *some* convenient verbal labels must often be given to them by general reviewers. All I did in my review was to categorize them with those "semantic blabs" which seem to be most commonly employed by discriminating psychological writers and teachers. If this is not scientifically sound, I should be happy to have Heinlein suggest a better procedure.

In paragraphs four and five, Heinlein seems to be scandalized by the results which would be obtained by converting r's into E's. While the facts of these paragraphs are substantially correct, his manner of relating them raises several implications which deviate from the truth:

1. Heinlein calls the E series "highly questionable," so that some readers might infer that there is something statistically wrong with it. This is of course not the case.

2. He emphasizes the fact that "the fifth category covers an extension of value greater than that of the first four categories." Naturally it does. As Hull (7), Bingham (1), Guilford (5), Garrett (4), Conrad and Martin (3), Peters and Van Voorhis (9) and others have pointed out, this is the distinguish-

ing feature of using *E*. Indeed, it is precisely this ultra-conservative interpretation which conversion into *E* gives to the lower degrees of *r* which recommends its used by test-makers who want to be certain that their personality questionnaires are adequate for individual diagnosis and prediction.

3. Heinlein again takes my term *unfair* out of context and—should I say *unfairly?*—over-emphasizes it in a manner I never intended.

4. The main implication of these two paragraphs seems to be that, while the verbal categories which I use in my article are too generous to those who have validated personality questionnaires in terms of *r*, the categories in terms of *E* (which I advocated for use at some *future* date) are too severe. Heinlein does not quite make clear, though, *why* the classification in terms of *E* is too severe, except to imply that it *looks* that way.

In his final paragraph, Heinlein points out that it is fantastic to "attribute cause-and-effect relationship or a determining, predictive function to an *r* as such, irrespective of the size of the *r*." Here he is apparently attacking all attempts to validate personality tests in terms of *r*. Now I hold no brief for test validation in terms of *r*. Nor do several other recent writers. Jackson and Ferguson (8), for example, advocate the use of analysis of variance for the calculation of test reliability and validity. Guilford (6) has recommended factor analysis. Brogden (2), on the other hand, has recently published a paper in which he substantially upholds the value of *r* as a direct measure of predictive efficiency, and contends that conversion into *E* is quite unnecessary. If Heinlein has still a different advocacy in this connection, I should be interested in hearing it. The main point is that in my article I merely noted that a good many experimenters—whether or not I, Heinlein, or anyone else likes the fact—*did* report test validations in terms of *r*. I was concerned, therefore, only with evaluating, or giving a convenient label to, these already reported *r*'s. I wholly agree with Heinlein that "an adequate proof of a pragmatic kind is sorely needed to demonstrate the efficiency of *r* as an index of validity for a specific type of personality questionnaire or test situation." My verbal classifications were certainly *not* designed to constitute that kind of required proof. Nor were they in the least intended to settle, once and for all, the many ticklish problems of measuring questionnaire validity. I am rather surprised that Heinlein should have ever thought this to be the case.

In sum, Heinlein's comment on my article on *The Validity of Personality Questionnaires* does not really call into question the general content or conclusions of my study but appears to boil down to (1) a questioning of the efficacy of *r* as an index of test validity, and (2) an attack upon the specific verbal categories that I attached to *r*'s actually obtained by test researchers. On the first of these points, I am inclined

largely to agree with his viewpoint and to welcome its airing. On the second one, I can only repeat that *some* verbal classifications of *r*'s must often be made for review purposes; that the ones I employed were based on long-term psychological usage; and that if Heinlein has any better suggestions to make, I shall certainly welcome them.

BIBLIOGRAPHY

1. BINGHAM, W. V. Reliability, validity and dependability. *J. appl. Psychol.*, 1932, **16**, 116–122.
2. BROGDEN, H. E. On the interpretation of the correlation coefficient as a measure of predictive efficiency. *J. educ. Psychol.*, 1946, **37**, 65–76.
3. CONRAD, H. S., & MARTIN, G. B. The index of forecasting efficiency, for the case of a "true" criterion. *J. exper. Educ.*, 1935, **4**, 231–244.
4. GARRETT, H. E. *Statistics in psychology and education.* (2nd Ed.) New York: Longmans, Green, 1940.
5. GUILFORD, J. P. *Psychometric methods.* New York: McGraw-Hill, 1936.
6. GUILFORD, J. P. New standards in test evaluation. *Amer. Psychologist*, 1946, **1**, 455.
7. HULL, C. L. *Aptitude testing.* Yonkers-on-Hudson, N. Y.: World Book Co., 1928.
8. JACKSON, R. W., & FERGUSON, G. A. *Studies on the reliability of tests.* Toronto: Department of Educational Research, University of Toronto, 1941.
9. PETERS, C. C., & VAN VOORHIS, W. R. *Statistical procedures and their mathematical bases.* New York: McGraw-Hill, 1940.

BOOK REVIEWS

MASSERMAN, JULES H. *Principles of dynamic psychiatry.* Philadelphia: W. B. Saunders, 1946. Pp. xix+322.

In 1943, J. H. Masserman published, under the title *Behavior and Neuroses* with the University of Chicago Press, an account of a long series of experiments which he had been conducting on cats. The experiments were designed to illustrate and elucidate behavior described largely under psychoanalytical categories. Masserman's new book, *The Principles of Dynamic Psychiatry*, extends the range of application of these experiments and undertakes a very ambitious program of formulation of principles and corollaries.

Masserman's system derives from the author's beginnings as a psychoanalyst and from an exposure to the Gestalt tradition plus wide reading in the field of general psychological theory, as well as his highly original line of experimentation. *The Principles of Dynamic Psychiatry* is divided into two parts of which the first (six chapters) covers the development of behavior theory in academic psychology and psychoanalysis. The second part (eight chapters) develops the author's system of biodynamics. He states his conception of the principles of behavior and finds illustrations from his experimental and clinical records. Masserman finds that there are four general principles.

1. Principle I reads: "Behavior is actuated by the physiologic needs of the organism and is directed toward the satisfaction of those needs." This is of course a very explicit adoption of the teleological point of view in its most naive and direct form. Masserman has no interest in investigating how a need can actuate behavior. He deals with end results and not means. This first teleological principle is supplemented by three others.

2. Principle II (experiential interpretation and adaptation) is: "Behavior is contingent upon and adaptive to the organism's *interpretation* of its total milieu, as based on its capacities and previous experiences."

3. Principle III (deviation and substitution) reads: "Behavior patterns become deviated and fragmented under stress and, when further frustrated, tend toward substitutive satisfactions."

4. The fourth principle (conflict) is: "When in a given milieu two or more motivations come into conflict in the sense that their accustomed consummatory patterns become incompatible, kinetic tension (anxiety) mounts and behavior becomes hesitant, vacillating, erratic, and poorly adaptive (neurotic) or excessively substitutive, symbolic, and regressive (psychotic)." (The statement of this fourth principle is a fair sample of the author's style.)

The four principles are each supplemented by numbers of corollaries. As may be inferred from the rather sketchy and loose statements of the principles given above, the corollaries consist of exceptions, relevant general remarks, or of elaborations of the principles. They are distinctly

not corollaries in the sense in which that word is used by mathematicians or logicians.

In an appendix, twenty-one pages are devoted to a detailed account of an illustrative psychoanalysis of a neurotic personality. Three other appendices include a short article on psychoanalytic formulations of the psychoses, an article on illustrative motion picture films prepared by the author and his associates, and a reprinting of an article on propaganda written by Masserman in 1942. Thirty pages are devoted to bibliography and forty to an extended glossary of psychiatric terms.

"Psychoanalytic theory today," Masserman says, " . . . is far from an accepted body of dogma; on the contrary, a great deal of it is fluid, ambiguous, unintegrated, and exceedingly polemic" [page 93]. Masserman's book should serve to reduce this confusion and advance the science of behavior. The interest of psychologists in the book will probably be directed less toward the principles and corollaries which do not bear very close scrutiny than toward the experiments with cats which are highly original in conception. Results of the cat experiments are interpreted as illustrations of principles and corollaries.

These experiments in general involved a cage, a means of administering punishment in the form of an air blast or shock and neutral stimuli and food rewards. They cover the simple acquisition of substitutive goals, fixation on part patterns, dominance, frustration, aggression, regression to earlier solutions, conflict-generating discriminations, experimental neurosis, and the effects on conflict of removal from the situation, forced solutions of conflict relief by re-training, inter-animal influence on neurosis, spontaneous working through to resolution of conflict, and the effects of drugs, particularly alcohol, on the behavior of animals in conflict situations. This is a rich experimental offering.

EDWIN R. GUTHRIE.

University of Washington.

ROGERS, C. R., & WALLEN, J. L. *Counseling with returned servicemen.* New York: McGraw-Hill, 1946. Pp. vii+159.

Although the words "servicemen" and "veteran" are frequently used and the cases illustrated are drawn from among returning servicemen, this book makes no attempt at a specific analysis of veterans' problems. Instead, it attempts to place such problems in the general framework of broad personality maladjustments to which the technique of nondirective counseling may be applied.

As a general introduction to the technique of nondirective counseling, the book is excellent. It is simply and clearly written, and the exposition of nondirective methods is orderly and reinforced by a wealth of carefully chosen quotations from actual counseling interviews. The last chapter contains practice exercises for would-be counselors, and

there is a selected bibliography of 23 titles in the appendix. Since the authors apparently envisage the practicing of nondirective counseling by readers whose instruction may be limited to the reading of this book ("Where skilled supervisory criticism is unobtainable definite profit may be obtained by a group of counselors studying together and analyzing and criticizing each others' interviews"), one wishes that the treatment of nondirective counseling were not so completely positive and enthusiastic, and that some mention had been made of negative instances or of the possible limitations of the nondirective technique in handling some types of problem.

WILLIAM A. HUNT.

Northwestern University.

JOHNSON, WENDELL. *People in quandaries.* New York: Harper, 1946. Pp. xiv+532.

In this book Dr. Johnson applies general semantics to the problems of personal and social adjustment. His thesis is (p. 45) " . . . that science clearly understood, can be used from moment to moment in everyday life, and that it provides a sound basis for warmly human and efficient living." Maladjustment, in Dr. Johnson's opinion, is essentially the result of *Idealism* which leads to *Frustration* and finally, to *Demoralization*. This IFD sequence, together with the myriad evaluations, labels, and abstractions which stem from it, leads to the conditions out of which many serious disorders of behavior develop.

The author presents a convincing case for his viewpoint, and he presents it in a clear and entertaining style rarely found in books on this subject. He spices his pages with a large number of examples which are both pointed and sprightly. For example, when discussing extensional devices, the author quotes a wag who reputedly said that there are two kinds of people in the world: those who always divide the people of the world into two groups and those who don't.

Most semanticists seem prone occasionally to yank Aristotle down from his niche in the halls of time and give him a thorough drubbing. Also, they like to indulge a similar mass action against the philosophers of any period. In this, Dr. Johnson is no exception. However, he devotes but a few pages to dispatching Aristotle and Aristotelian logic, and his potshots at philosophers are few and mild. In contrast, he treats the members of the medical profession with exaggerated respect, almost with awe. He might have presented some excellent examples of how the members of some medical groups created problems of a semantic nature by hanging labels as *psychoneurotic* or *NP* on many thousands of servicemen.

A minor fault in the book is found in the author's occasional tendency to cast aside logical restraint when elaborating a point. On page 280 he

writes, "Einstein has never performed a laboratory experiment; that may well be a major part of the reason for his tremendous scientific achievements."

But such lapses are quite rare. The book represents a significant contribution in an area which has hitherto been inadequately treated. While it is really a book for everyone, it has a special appeal for those who deal professionally with problems of adjustment. Clinical psychologists will find particular value in the chapter "In Other People's Quandaries" and in the semantic exercises and research discussion at the end of the book. As a text or as reference reading, *People in Quandaries* should fit in well with psychology courses which deal with behavior problems.

IRWIN AUGUST BERG.

University of Illinois.

CARP, BERNARD. *A Study of the influence of certain personal factors on a speech judgment.* New Rochelle, New York: The Little Print, 1945. Pp. viii+122.

The field of speech and speech education has expanded rapidly in recent years. In the selection of teachers, radio announcers and speakers, stage and screen performers, telephone operators, etc., as well as in placement work in the schools, the need for improved methods of speech testing has become acute.

This study attempts to determine whether audible speech can be tested and rated objectively. It also describes in detail the methods used in constructing, administering, scoring and weighting an original battery of speech tests called Speech Appraisal Forms.

Speech tests were administered to 25 male college seniors and graduate students whose test performances were rated on Speech Appraisal Forms by 24 specially trained judges. The resulting data were treated statistically by means of the Latin Square modification of the "Analysis of Variance" technique. The author recognizes the limitations of this method and recommends further experimentation and more precise procedures. He concludes, however, that such factors as dress, appearance, poise, etc., "need not significantly affect a judge's rating of audible speech."

This work should be of particular interest to those concerned with speech testing of either children or adults. It will be of value, also, to those interested in conducting further research in this important field.

ST. CLAIR A. SWITZER.

Miami University.

MORGAN, JOHN. J. B. *How to keep a sound mind.* (Rev. Ed. of *Keeping a sound mind.*) New York: Macmillan, 1946. Pp. vii+404.

The point of view expressed in this book is identical with that of the

first (1934) edition, i.e., that mental health depends upon the development of correct mental habits. Ten of the 14 chapters are essentially the same as ones in the older book, although the titles are not always the same. The materials have been rearranged and improved, new sections have been added and others deleted. The chapters on crime, self-confidence, exaggeration of defects, and how to strive toward desirable goals have been replaced with chapters dealing with a wholesome pattern of living, developing social poise and security, and "how to be happily maladjusted." Some of the materials in the omitted chapters appear at various places in the new book.

The brief bibliography at the end of the earlier edition has been removed and a list of references included for each chapter. Considerably fewer than half of the titles, however, bear dates later than 1934. The number of study questions has been greatly reduced. There are no tables, graphs, or illustrations.

The style is interesting, with many case studies and other valuable anecdotal material. There is, however, a lack of documentation and of references to experimental studies. Occasionally statements are made which are questionable and should be supported by data. For example (p. 230), "The probabilities that a man will remain in an occupation depend (according to statistical studies) much more on the financial and domestic responsibilities which he must meet than upon his success in or fitness for his job." No evidence is cited.

The book, offered as a basal textbook for college classes in mental hygiene, attempts to " . . . [put] in understandable form the basic principles involved in the preservation of one's own mental health." This purpose is well achieved, although the book perhaps will be more useful as supplementary material than as a basic text in mental hygiene, especially with students who have had previous work in psychology.

CLAUDE M. DILLINGER.

Illinois State Normal University.

STERN, EDITH M. *The attendant's guide.* New York: The Commonwealth Fund, 1945. Pp. xiv+104.

Recent "revelations" of alleged mistreatment of psychopathic hospital patients has naturally raised questions about, and brought charges against, attendants employed in such institutions. Stern's little book is a timely instrument for improving the training of attendants.

Divided into three parts, it treats respectively of procedures demanded by patients' general characteristics, particular kinds of patients, and of the attendant and his job. The first part has to do with hospital routine. Part Two classifies patients by dominant symptoms (from the view point of ward care), for example, patients "who do too much," are "not fit to be seen," "convalescents," and other similar groups. The final part is a frank talk about the attendant's responsibilities and his

present limited prestige status. The advantages of the occupation are listed as valuable experience prefatory to more advanced levels of training, and being in on the ground floor of a coming skilled craft.

Throughout the style is direct and non-technical. Chapters are replete with specific suggestions for meeting practical problems. Mrs. Stern's book will make a definitely favorable contribution to the goal it professes, namely the improvement of attendants' skill.

STANLEY S. MARZOLF.

Illinois State Normal University.

HUNTINGTON, ELLSWORTH. *Mainsprings of civilization.* New York: John Wiley, 1945. Pp. xii+660.

Professor Huntington's general thesis concerning weather and climate in relation to human affairs is widely known through his many publications. This book attempts an integration of much of his work to date.

The mainsprings are: A "basic evolutionary force," the innate qualities of the human organism, the physical environment, and the social environment or culture. Although discussed more specifically in the author's earlier writings, these basic concepts in the present context are not concisely expressed. For example, the "great evolutionary force which permeates all nature" is implicitly accepted as a kind of push from the rear operating in the history of life. Although told that "for thousands of years civilization has been advancing along certain definite lines," the reader is left to infer from context that these "definite lines" probably have to do with the development of larger and larger single units of social organization and with the "betterment" of living generally. Again, the innate qualities of the organism are referred to in very general terms, sometimes almost as a primary good, insuring human progress.

The writing is enthusiastic and vigorous and will undoubtedly appeal to those who seek to establish correlations among the several disciplines addressed to the study of human behavior. The book will be read with some skepticism by those who are well acquainted with the "ifs," "ands" and "buts" pertaining to the facts established within any one discipline.

Professor Huntington disavows the theory of innate race differences; he states that all peoples or stocks contain qualities which will enable them to adapt and to develop culturally, but that these qualities develop or fail to develop depending upon the presence or absence of appropriate environmental conditions. An ingenious comparison of Newfoundland and Icelandic cultures illustrates this interplay of organic quality and environing circumstances.

Social psychologists will be interested in the treatment of character and inheritance, the place of "kiths" ("a group of people relatively

homogeneous in language and culture and freely intermarrying with one another") in history, and the significance of selective migration and environment in the development of kiths. Likewise, the consideration of weather in relation to mental activity and behavior, diet and national character, and cycles in human activity deserve attention.

Psychologists will find Huntington's criterion of mental activity somewhat limited: "A good measure of intellectual activity on a large scale is the circulation of books by libraries, especially ordinary city libraries. People read serious books more frequently when their minds are active than when they are inert" (p. 344). Armed with library circulation figures from a number of American cities, the author demonstrates a seasonal and a weather linkage with reading and advances a theory of heightened mental activity in relation to increased atmospheric ozone.

In only a few instances, chiefly citations from literature, are correlation coefficients presented. At one point (p. 230), the author refers to r^2 as giving the "per cent of the resemblance between two variables due to a common cause"; it might be more accurate to speak of "common elements." This apparent lack of distinction between concomitance and causation appears at a number of points throughout the book. Causal relationships are freely deduced or inferred from observed associations. In practically no instance is the degree of the association established in terms of a correlation coefficient, or are the observed associations submitted to a test of significance.

The author accepts the concepts of multiple causation and of multiple correlation, but he experiences great difficulty in applying them to his data by verbal, descriptive methods. There is an almost complete absence of the concept of reliability of a statistic. When percentages are compared, differences which fit a "trend," no matter how slight, are accepted as "real." Occasionally, the author will advance rational explanations for slight reversals in a trend.

The chief problem which psychologists will have in evaluating this work is that of the quite different framework within which the author operates. His interest in human behavior is oriented toward man in the mass rather than toward man, the individual. He concentrates on trend rather than on variation. Huntington's procedures probably do reveal facts which analysis, by segmenting data, tends to overlook. Roughly analogous situations occur in the relation of range of measurement to magnitude of correlation, and in the use of the correlation coefficient for actuarial prediction as contrasted with prediction of the single case.

But in spite of the author's apparent unfamiliarity with psychological methods and findings, this study has considerable value for psychologists and their work. It reveals a body of material of undoubted significance for human life which can greatly complicate the psycholo-

gist's effort to account for variance in, and to predict the course of behavior. It affords a mass of ideas which could well command the research energies of a corps of social psychologists for a lifetime, translating the ideas into hypotheses capable of experimental test.

DALE B. HARRIS.

University of Minnesota.

SMITH, BRUCE LANNES, LASSWELL, HAROLD D., & CASEY, RALPH D. *Propaganda, communication, and public opinion; a comprehensive reference guide.* Princeton: Princeton Univ. Press, 1946. Pp. v+435.

This book represents a continuation of *Propaganda and promotional activities; an annotated bibliography*, also written by the present authors, and published in 1935. It lists titles of and briefly comments on nearly 3,000 selected books, periodicals, and articles, most of which appeared between mid-1934 and March, 1943. In addition to this annotated bibliography, which fills approximately two-thirds of the book, there are certain other features of general interest.

Four chapters at the beginning cover important background material: channels of communication, political communication specialists, description of contents of communications, and description of effects of communications. In his section on communication channels, Casey briefly reviews the historical development of American communications as influenced by the rise of democracy, industrialization, and urbanization. The economic aspects of communication are discussed, with special reference to costs of and relationships between radio broadcasting and newspaper publishing.

Smith follows with a discussion of the major political communication experts of our times. Information is given in tabular form concerning the heads and propaganda ministers during World War II of some of the larger nations. Their occupational origins are noted and data are presented concerning their fathers, and their own careers. Further tables give such facts about them as estimated incomes of fathers, childhood exposure to authoritative symbols of society, formal education, socio-economic status during first decade of employment, and ages.

In the third and fourth introductory essays Lasswell attacks the problems of describing the contents and determining the effects of communications. Analysis of contents can be done by a listing and comparison of themes, by counting of socially optimistic and pessimistic expressions in speeches, and noting values expressed in motion picture characterizations. It is desirable to base content analysis on insight into subject response. Effects may be judged by votes in free elections, actions by public officials, by people in direct contact with the movements involved, etc.

The bibliography itself is arranged in sections covering (1) strategy and technique, (2) promoting groups, (3) response to be elicited, (4) symbols utilized, (5) channels, (6) measurement, and (7) control and censorship. Entries are arranged alphabetically by titles within sections. One hundred and fifty outstanding titles are starred in the listings and commented on relatively more extensively. This sectioning and alphabetization is supplemented by a comprehensive combined subject and author index.

As in the case of many books by several authors, the individual sections are not too well integrated, although in themselves interesting and informative. Annotations do not seem to follow any particularly systematic pattern, covering in various combinations the actual contents, their significance, their value, and the backgrounds of the authors. This approach sometimes leads to weak notes, but on the whole the annotation is quite adequate for most purposes. The 42-page author and subject index should add greatly to the usefulness of the bibliography.

This book is a reference work of great obvious value, indispensable to any adequate library.

R. B. AMMONS.

University of Denver.

DEWEY, JOHN. *Problems of men.* New York: Philosophical Library, 1946. Pp. iv+424.

Except for a prefatory note and an introduction of 18 pages, this book consists wholly of a reprinting of articles published during the past 12 years in 14 different journals and magazines, together with one paper originally published, as stated in the prefatory note, "half a century ago" (exact time and place not indicated in the book, but actually in *Decennial Publications of the University of Chicago*, 1903).

A reviewer in *The New York Times* (June 9, 1946) has called this "a mighty book." This may be true in the sense that its author is a mighty man of influence, through more than half a century, on psychology, education, and philosophy. His influence was felt earliest on psychology (functionalism) through his 1886 book, *Psychology*, and through his 1896 article, "The Reflex Arc Concept in Psychology." Since the turn of the century his influence has been greatest on educational theory and practice ("progressive education") and on philosophy (instrumentalism). Whatever book he writes will command the respect of thoughtful readers. What E. G. Boring wrote of Dewey in 1929 is equally true today,—"While he has had in the last twenty-five years little effect upon psychology proper, he has led a very influential life in its effects upon intellectual America, expounding repeatedly the problems of human nature" (*A History of Experimental Psychology*, 542).

Except for the magic of the author's name, however, this book is not

great *as a book*, since it was originally written, not as a book, but as individual articles, each complete in itself, for a wide range of journals and magazines from *The Journal of Philosophy* to *The Rotarian*. Few readers except reviewers and devotees of Dewey's philosophy in its entirety will have the persistence to read the whole book. The chapters on education will be of interest to the general reader, dealing as they do with a vigorous defense of democracy and the experimental method against all forms of external authority, whether these are based on classical tradition, medieval supernaturalism, or present-day demands for "a moratorium on science." On the other hand, all the chapters of Part III, except one, are from *The Journal of Philosophy;* and they consist largely of technical philosophical arguments of interest mostly to other technical philosophers. The uncut pages typically found in numbers of that journal in the average college library bear witness to this statement.

One reviewer has spoken of Dewey's "inimitably cluttered prose" (*Time*, June 26, 1946). While this is a true enough characterization in general, there are instances in this book of chapters, like the one on James Marsh and the first of the two on William James, so well written as to make applicable to Dewey himself what he says of Bertrand Russell,—"His lucidity and felicity of expression are ever the despair of lesser writers, and in . . . [these chapters] he has almost surpassed himself" (p. 171).

WESLEY R. WELLS.

Syracuse University.

RAND, W., SWEENY, M. E., & VINCENT, E. L. *Growth and development of the young child.* Philadelphia: W. B. Saunders, 1946. Pp. vii+481.

The fourth edition of this elementary text of child growth and development appears six years after the third revision. The subject matter is almost completely reorganized, but much of the old content remains essentially unchanged. The new edition is arranged in topical sections—physical, intellectual, social and emotional, etc.—whereas, the earlier edition considered all of these topics in sections devoted to different chronological periods of development. While the more recent treatment provides a more logical and continuous approach to child development, it may make the book more cumbersome and inconvenient for parents who wish to read as their children develop.

The sections of this new edition devoted to emotional, intellectual and social development have been expanded and a chapter on current concepts of growth and development has been added. The chapter on current concepts reflects a significant trend in child care—a trend away from Watson's dicta toward the contemporary philosophy of self-regulation of diet and development as proposed in recent publications by Gesell and Ilg, Ribble, the Aldriches, *et al.* An example of this shift

in emphasis is the authors' discussion of thumbsucking. In the third edition they state, "Mechanical restraints are sometimes helpful if used before the habit has become so important to the child, or if used with older children as a reminder when they themselves have decided to conquer the habit." In the recent revision one finds "The older recommendations for the use of mechanical restraints, of rewards and punishments and other adult imposed devices are now considered not only poor practice, but in all probability risky to the future well-being of the child." References to the recent publications of Gesell and Ilg, and Ribble are copious.

Although the new edition contains 417 references, as contrasted with only 221 in the third edition, one still finds many aspects of psychological development ignored, sketchily treated, or inadequately supported with available, published research data. In some cases, definite rules are laid down for child guidance without any reference to research findings. This, in many cases, appears to be an effort to extend a hand of aid and comfort to parents who want *something definite*—correct or incorrect.

Since this revision retains much of the previously published material and includes a considerable amount of recent research, it should be accepted and appreciated by those who have found the earlier editions useful. In the reviewer's opinion this edition should be a helpful source book for a highly selected group of parents and a better-than-average textbook for an elementary course in child care and training.

GEORGE G. THOMPSON.

Syracuse University.

BEAUMONT, HENRY. *The psychology of personnel.* New York: Longmans, Green, 1945. Pp. xiii+306.

BEAUMONT, HENRY. *Psychology applied to personnel.* New York: Longmans, Green, 1946. Pp. viii + 167.

The Psychology of Personnel is intended for employers and as a text in a general course for college students. The level of the text is such that a previous course in psychology would not be necessary as a prerequisite. *Psychology Applied to Personnel* is a work-book to accompany the text.

The Psychology of Personnel contains the following 11 chapters: I. Understanding Employees, II. Analyzing Jobs, III. Selecting Employees, IV. Training Employees, V. Working Conditions, VI. The Workers' Health, VII. Promoting Safety, VIII. Supervision, IX. Merit Rating, X. Providing Incentives, XI. Occupational Adjustment. Citations from 59 industrial and other organizations, which are an important contribution of the book, give examples of personnel procedures. Thirteen organizations are cited with reference only to Training Em-

ployees. There is no author index, and in fact no author or almost none is cited by name in the entire book. There are no tables, no figures, and practically no psychological statistics. This type of presentation of the material which omits the names of investigators and statistics will probably appeal to the average reader.

As is somewhat implied by its title, the treatment of *The Psychology of Personnel* is intermediate between books on industrial psychology and personnel administration. The absence of statistical results and freedom from detailed reports of investigations makes it appear closer to the usual book on personnel administration than to the usual book on industrial psychology.

Beaumont has succeeded in writing a general book that is probably simple enough not to intimidate anyone. Almost all of the available conclusions of psychological applications to personnel are presented. On the other hand there is little perspective shown in pointing out major emphases. There are an unusually large number of topics included for a book of this size. Besides the topics expected from the chapter headings there are a good many surprising ones as for example a brief discussion of Planned Parenthood (p. 165). The simplicity of the book occasionally spills over into elaborating the obvious as "In all plants, offices, and stores where women are employed there is a need for separate sanitary facilities." And with reference to evaluating experience, an entire page is devoted to pointing out that whether the experience of an auto mechanic in Detroit will be indicative of success in Tulsa depends on level of performance, pay level, hours of work, location of plant, and job environment (p. 60).

In addition to presenting a great many examples of personnel practices in American industry and business, considerable emphasis is placed on Army practices and on the problems of servicemen returning to civilian employment.

The two fullest topics are those of training and selecting supervisors by tryout.

Psychology Applied to Personnel is made up of two parts, Part I: Personnel Statistics, and Part II: Notes, References, Questions, and Applications.

Part I: Personnel Statistics, gives the method and blank work sheets for calculating frequency distribution, measures of central tendency, range, standard deviation, group and individual comparisons, significance of traits, and product-moment correlation. A set of actual data on 94 cases from a company is given. These data are used in the illustrative problems and are to be used in assignments. These assignments are interesting. Uncharacteristic but nevertheless unfortunate is a slip in explaining correlations (p. 36). The explanation assumes that the relative importance of coefficients of correlation is of the same

magnitude as the size of the coefficients rather than the square of the coefficients. This error is not cleared up under *The significance of correlation*, when referring to a coefficient of −.09, " . . . this relationship is not a close one so that there will be many exceptions to this general rule." Wonderful understatement!

Part II of *Psychology Applied to Personnel* has chapter headings identical with those of *The Psychology of Personnel.* Twenty-five good true-false questions are presented for each of the 11 chapters of the text.

A large number of up-to-date references are given with respect to each major topic of each chapter. These references are from a variety of sources, industrial magazines, personnel journals, advertising or semi-advertising publications by companies, government publications, and the psychological literature. The references should be particularly useful to teachers. For the average employer or other student there are too many references and insufficient annotation to be of greatest usefulness. A deficiency with respect to a number of references to government sources is that the references are not satisfactorily definite, as for example, "The Women's Bureau of the United States Department of Labor has available several publications dealing with proper provisions for the employment of *women* in industry" (76).

All in all these two books by Beaumont are a definite contribution in presenting personnel psychology at the intended elementary level.

THOMAS W. HARRELL.

University of Illinois.

MAIER, NORMAN R. F. *Psychology in industry.* New York: Houghton Mifflin, 1946. Pp. xvi+463.

Reports during the last decade describing studies of the effect of the social milieu in industry are reflected in this new *Psychology in Industry.* Although most of the topics covered are those found in the conventional industrial psychology text, the emphasis and amount of space allotted to each topic is considerably different. While this book is not a manual of advanced techniques, the discussion is by no means shallow.

About one-quarter of the book considers general principles and causes of behavior, and the development of attitudes and morale. The problems of the worker on the job and his relationships to it, his fellow workers, and to management are discussed in detail. Maier feels that one must understand human behavior as it exists rather than to blame and punish. Man's behavior is a product of multiple forces and these must be changed before behavior changes. "Causation implies that a given individual in a given situation must do as he does." Management should seek causes rather than treat symptoms. Reference to the Western Electric studies shows the effect of social groupings, attitudes,

and morale upon performance. Since the background for the development of attitudes in the cases of management and labor is different, and since there is no "right," mutual understanding of the other's point of view is necessary.

The section on individual differences, ratings, and the use of tests is elementary and general. Where other texts become expansive and specific, Maier deliberately curtails. Practically no specific tests are mentioned although the reader is referred to other sources. This section of the book is the weakest and most poorly organized. Even a non-technical review should emphasize the extreme importance of criteria. The consideration of job analysis (and its relationship to selection of adequate criteria) should precede material on selective and evaluative instruments.

Fatigue, time-motion analysis, accidents, and the working environment (illumination, atmosphere, noise) receive adequate attention. The author comments on the techniques employed by the industrial engineer, but his emphasis is on changes in the behavior of workers brought about by improved conditions. Training and motivation get standard treatment. Perhaps the topic of motivation should have been included in the earlier chapters concerned with causation in behavior. Labor turnover is interpreted as a symptom of dissatisfaction and management is advised to ferret out its causes. The final chapter reviews briefly, but perhaps too repetitiously, the material covered in the preceding chapters with special emphasis upon its use by the supervisor, the industrial counselor, and higher management.

A number of errors or questionable statements can be found, but probably not more than tend to creep into any text. For example, there is an implication that any test that will correlate .30 with a criterion is good, and while one grants that predictions will be better than chance, the implications of such a low validity coefficient should have been clarified. In fact, the author once implies that any test is better than no test. In another place one finds the statement, "Intelligence tests . . . are aptitude tests in the sense that education and experience have little or no effect on the score." One statement says that jobs requiring dexterity and mechanical operations show *no* relationship with intelligence, but the following sentence indicates that lower levels of intelligence are sometimes more satisfactory on these jobs. In reference to time-motion economy, the view that, "The right and left halves of a man's body are mirror images of each other," disregards the functional inequality of opposite sides of the body.

The physical appearance of the book is good but appropriate illustrations might make for higher interest, particularly in view of its intended audience. The few references made are found at the bottoms of pages which allows easy reading. A bibliography by chapters is found in the appendix.

If used as a text, supplementary readings would be desirable. Certainly space allotments do not actually reflect the amount of experimental information available in each area. However, the change in emphasis toward the social and individual aspects of work and the variety of applications suggested make refreshing and interesting reading for the student of industrial relations.

LESTER P. GUEST.

The Pennsylvania State College.

HARRIMAN, P. L. (Ed.) *Twentieth century psychology.* New York: Philosophical Library, 1946. Pp. xiii+712.

This book does not fulfill the expectations implied by its attractive title. The reader who expects a survey of contemporary psychology will be disappointed. It is a collection of thirty-nine articles by an assortment of writers, most of them well-known to psychologists, and some of them heavyweight authorities. But twenty-five of the articles are reprints from current periodicals, the majority easily accessible to professional psychologists. Periodicals represented more than once include the *Journal of Social Psychology* and the *Journal of Abnormal and Social Psychology*, with four articles each, the *Journal of General Psychology* with three, and the *Psychological Review* with two and the larger part of another. Thus, many of the contributions will not be new to the profession.

According to the editor, the volume was prepared for the general reader, but it is uncertain what general reader he had in mind. The psychologically untrained reader will find the technical jargon too much for him. The student coming up in the field will find much better volumes for parallel reading in his course work, until he is ready to go directly to the periodical literature. About the only general reader I can think of for whom the volume might have been designed is the specialist in related fields of knowledge, e.g., philosophy, the social and natural sciences, etc. Such a reader is apt to get as much impression of what is wrong with psychology as of what psychologists are contributing of use and value to him.

One thing which would appear to be wrong is that psychologists seem to write with a heaviness of touch that is well-nigh depressing. Let me cite only two of many possible examples. "How may differences in subgroup structure, group stratification, and potency of ego-centered and group-centered goals be utilized as criteria for predicting the social resultants of different group atmospheres?" ask Lewin, *et al.* (200). "We may say that the disturbance of the tentatively established movement-stereotypy at this critical point disordered the manner in which the focal stimulus-patterns were encountered just before blocking occurred in the given blind, thereby interfering with expansion of the nuclear blind-conditioning process in the segment," says Schnierla

(291–292). The general reader who wishes to know more about what is being discussed in passages like these may refer to previous publications of the writers. He may, in short, begin specializing in psychology and, in fact, in their brand of psychology.

The book may also be censured for a carelessness of production that is hardly excusable on the ground of wartime difficulties. There are many typographical errors, occasional misplaced footnotes, and one complete failure of correspondence between illustrative figure and text (421–424). The text refers in capitals to A, B, C, D, E, F, G and H, and to shaded areas HEG and GFH. The figure, if such it be, though it isn't so labeled, has the lower case letters a, b, c, d, e, f and g scattered over the page, and no shaded area. Finally, there is a two-page index of one column to the page with a total of 70 entries for the more than 700 pages of material.

A third weakness is an emphasis, unbecoming of a field struggling to assume a proper position in the natural sciences, on the armchair method. One may have no objection to the use of intuition and insight whenever possible, but if Klein, who writes in defense of this method, wishes to see some prime examples of what it would contribute to the advancement of knowledge, he has only to read the closely following articles by Maslow and by Brunswick. The first refers to some percentages of strength of motivation (40) that are plainly imaginative, and the second is abracadabra.

Lest this review suggest that there is nothing worthy of publication (or republication) in the book, I might mention that I enjoyed many of the original articles, including most of the review of his research by Schnierla, the article on conditioning by Harris, Brandt's survey of his ocular photography (although a little more modesty might be in order in the descriptions of his apparatus), and, among the reprints, Harrower-Erickson's outline of Rorschach methods. Also, the article contributed by the editor is presented simply enough to indicate that he, at least, knew what general reader he had in mind. But these seem to be a meagre return for the heavy expenditure of time necessary to go through the whole volume.

GEO. M. PETERSON.

University of New Mexico.

BOOKS AND MATERIALS RECEIVED

BENEDEK, THERESE. *Insight and personality adjustment.* New York: Ronald Press Co., 1946. Pp. xi+307.

BLACK, IRMA S. *Off to a good start.* New York: Harcourt, Brace, 1946. Pp. xii+256.

BLANKENSHIP, A. B. (Ed.). *How to conduct consumer and opinion research. The sampling survey in operation.* New York: Harper, 1946. Pp. xi+314.

CLEETON, G. U., & MASON, C. W. *Executive ability. Its discovery and development.* Yellow Springs, Ohio: Antioch Press, 1946. Pp. 540.

COOKE, E. D. *All but me and thee: psychiatry at the fox-hole level.* Washington: Infantry Journal, 1946. Pp. 215.

DAVIS, F. B. *Item-analysis data. Their computation, interpretation, and use in test construction.* Harvard Educ. Papers No. 2. Cambridge: Graduate School of Education, Harvard Univ., 1946. Pp. v+42.

DE GRUCHY, CLARE. *Creative old age.* San Francisco: Old Age Counseling Center, 1946. Pp. iv+143.

DUNLAP, K. *Religion—its functions in human life.* New York: McGraw-Hill, 1946. Pp. xi+362.

FALES, W. *Wisdom and responsibility.* Princeton: Princeton Univ. Press, 1946. Pp. 166.

FRIEDMAN, BERTHA B. *Foundations of the measurement of values.* Teachers College, Columbia Univ., Contrib. to Educ., No. 914. New York: Bureau of Publications, Teachers College, Columbia Univ., 1946. Pp. viii+227.

GARRISON, K. C. *The psychology of adolescence.* (3rd Ed.) New York: Prentice-Hall, 1946. Pp. xv+355.

GRAY, J. S., *et al. Psychology in human affairs.* New York: McGraw-Hill, 1946. Pp. viii+646.

HARRIMAN, P. L. (Ed.). *Encyclopedia of psychology.* New York: Philosophical Library, 1946. Pp. vii+897.

HARTLEY, RUTH E. *Sociality in pre-adolescent boys.* Teachers College, Columbia Univ., Contr. to Educ., No. 918. New York: Bureau of Publications, Teachers College, Columbia Univ., 1946. Pp. viii+99.

KINGSLEY, H. L. *The nature and conditions of learning.* New York: Prentice-Hall, 1946. Pp. xvi+579.

LANDIS, C., & BOLLES, M. MARJORIE. *Textbook of abnormal psychology.*

New York: Macmillan, 1946. Pp. xii+576.

LEWIS, CLAUDIA. *Children of the Cumberland.* New York: Columbia Univ. Press, 1946. Pp. ix+217.

MCFARLAND, R. A. *Human factors in air transport design.* New York: McGraw-Hill, 1946. Pp. xix+670.

MAURER, KATHERINE M. *Intellectual status at maturity as a criterion for selecting items in preschool tests.* Minneapolis: Univ. of Minn. Press, 1946. Pp. v+166.

MONCRIEFF, R. W. *The chemical senses.* New York: John Wiley, 1946. Pp. vii+424.

PORTERFIELD, A. L. *Youth in trouble.* Ft. Worth: The Leo Potisham Foundation, 1946. Pp. 132.

RICHARDS, T. W. *Modern clinical psychology.* New York: McGraw-Hill, 1946. Pp. xi+331.

SHAKOW, D. *The nature of deterioration in schizophrenic conditions.* Nervous and Mental Disease Monographs, No. 70. New York: Coolidge Foundation, 1946. Pp. vii+88.

THORNTON, N. *Problems in abnormal behavior.* Philadelphia: Blakiston, 1946. Pp. x+244.

WORTIS, H., SILLMAN, L. R., & HALPERN, FLORENCE. *Studies of compulsive drinkers.* New Haven: Hillhouse Press, 1946. Pp. 90.

Vol. 44, No. 5 September, 1947

Psychological Bulletin

Since the beginning of the century there has been a volume of work studying spontaneous activity of animals, especially the rat and monkey. The best review article is that of Shirley (97) in the *Psychological Bulletin* in 1929. Since that time the subject has not been reviewed comprehensively. Somewhat limited reviews appear in Munn (67), Morgan (66), and Gray (31). Other and even more circumscribed reviews include Richter (73, 74) reporting work done under his direction; Hoskins (40), describing some of the relationships between endocrines and activity; and Kreezer's (52) summary of methods for measuring activity in the rat. A review of diurnal rhythms by Welsh (119) in 1938 discusses much material not directly related to activity. Mettler (64) has reviewed and summarized studies on the effects of striatal injury in 1942.

This article does not attempt to cover work done prior to Shirley's review in THIS JOURNAL. Several of the most important references to work done prior to 1929 are included, but the emphasis has been almost entirely on later material.

THE CONCEPT AND MEASUREMENT OF SPONTANEOUS ACTIVITY

There are two points the reader should keep in mind as he proceeds through the paper. The first is a methodological issue. Much of the research to be reviewed depends on a general concept of spontaneous activity without regard to how the activity is measured. It will become evident in the course of the review that our concept of activity must be tied to the measure of it which we have used, for the results one gets with one measure of activity may be entirely reversed when a different measure is used. The second closely related point is a matter of terminology. Since the largest amount of work has used animals running inside a drum, it will simplify things if the term *activity*, without any qualification, always refers to running activity, not to other measures

of activity. Measures other than those in an activity-drum will always be clearly distinguished.

Method and Apparatus

Running Drums. Animals and human beings indulge in spontaneous activity. This observation has been quantified in many ways. The animal most frequently used in experimentation is the rat, whose activity is usually measured in what has been called an activity cage, but will henceforth be referred to as a drum or running drum. This device was first used by Stewart (110) and has been most adequately described by Slonaker (101, 102). It usually consists of two 10-13 inch circular boards mounted on a shaft and separated by a sheet of mesh wound around their periphery (86, 94). The rat runs inside the freely rotating drum, and a counter is attached to record the number of revolutions. Unfortunately, the usual system of measurement has shown only total activity, not activity as a function of time.

Recognizing this inadequacy, Skinner (100) used a Harvard work adder in conjunction with a kymograph to get a summative record whose slope is a constant measure of activity.

The drum has almost as many variations as there have been experimenters in activity. Stewart's 20-inch diameter drum and the 26-inch diameter drum used by Park and Woods (71) represent one extreme, while Shirley (94) used a 10-inch diameter. Results reported in terms of number of revolutions are obviously not comparable when the diameter of the drums is not the same. Furthermore, equating the running by expressing it in distance traversed is of questionable validity in view of Farris' statement that rats in larger wheels run farther than those in smaller wheels (24).

Depending upon the experiment, the rat may live entirely within the drum (94, 95, 101), have a separate living cage, or use supplemental diffuse activity cages (71). Since Richter (73) has shown that the number of revolutions of the drum is reduced when the rat has a choice of several things to do, the results of different experimenters may not be comparable.

The revolving drum has been the most extensively used laboratory instrument in investigating activity. Its physical variables have been discussed by Skinner (100) and Lacey (53). The reliability of the measures obtained is remarkable—Shirley (94) reports a rank-order correlations of .97 for five-day totals of activity, and a split-half r of .90. Beach's (4) figures are even higher (.98). Unfortunately a basic assumption in these results involves the equivalence of the measures. Lacey

(53) raises the very justifiable criticism that the measure may be showing only the consistency of the different drums. It is significant to note that in one case in which the animals were changed from one cage to another, the correlation reported was .80 (113). There are wide individual differences even between litter mates in normal rats with respect to running, some rats running 200 revolutions per day and others 20,000. The pattern of running is set up by the tenth day or not at all. After this time the individual differences are relatively constant.

The running drum has been used to indicate tension or motivation in the rat. Thus, Durrant (20) and Slonaker (106) have correlated running with sex drives. Geier and Tolman (28, 29) have used running behavior to indicate increase in tension in the rat.

Dorcus (16) devised a cage which moved slowly toward a goal object when the rat ran inside of it.

Tambour- or Spring-Mounted Cages. Another apparatus for measuring activity is that first used by Syzmanski (114, 115) which consisted of a spring-mounted cage attached to a lever recording system. The disadvantage of lack of damping has been somewhat overcome by tambour-mounted activity cages (73, 109). The three supporting tambours are joined to one tube and record every movement on a kymograph. Both these methods produce records according to time, but records which are difficult to treat quantitatively because no ready means of determining the total activity is available.

Hunt and Schlosberg (42, 43) counted the number of 5-minute active periods occurring in such a cage over varying intervals of time, and Irwin (44) recorded the number of active seconds per minute in newborn children. Wilbur (121) used a spring mounted cage connected to a Harvard work adder to obtain a summative record (of the activity of chicks) which is much easier to interpret.

Smith (109), measuring audiogenic and electrogenic convulsive activity, supported a cage from four pneumographs or by one large flexible hydron bellows. Oscillation could be reduced by means of a small vent.

Other animals have been used with appropriately modified cages to record activity. Monkeys have been fastened by a nine-inch chain to a 2.5-inch rod, so that movement caused the rod to advance a counter (84). A monkey-sized pneumatically-mounted activity cage has been used by Kennard, Massimy and Chevallier (51, 62).

Other Automatic Methods of Recording Activity. Another measure of activity has been suggested, incorporating a tilting box (7) in which the movement of the rat from one end of the box to the other advanced a counter. Claiming that the tilting motion would interfere with accurate

measurement of rat's activity, Siegel (**98**) utilized the animal's motion from one end of a 22×6 inch box to the other to break a photoelectric relay and thus advance a counter.

A horizontal turntable for exercising rats has also been used to record activity (**21**, **22**). Since the distance the rat runs depends upon his proximity to the center of the turntable, it is probable that this device will not be popular in controlled experiments.

Curtis (**15**), working under Liddell, reports the use of a pedometer to record activity of the sheep and the pig. Head-shaking in chickens has also been reported by Levy (**58**).

Observational Method of Quantifying Activity. An observational means of recording activity has been used by Hall (**32**), Beach (**3**), and Fredericson (**25**). Hall recorded the distance traversed by rats in a round open field eight feet in diameter. Beach noted which of 36 squares a rat entered upon in the ten minutes it was free in an area three feet square. Fredericson observed six classes of behavior indulged in by rats in a field two feet square.

Spontaneous Activity as a Behavior Category

Now let us take a moment to see what the methods just reviewed have to do with the concept of activity. Most of the authors tend to lump all manifestations of activity together and to pin one label on all of them—*activity*. This failure to distinguish types of activity in terms of its measure leads to a false concept of activity, for what data we have point to more than one type, or at least more than one aspect of activity.

There are, for example, wide individual differences in the running activity of rats but not nearly such wide differences in restless cage activity. Tainter (**116**) found that caffeine, metrazol, and picrotoxin had no effect on running but did increase behavior measured in a diffuse activity cage. Hunt and Schlosberg (**43**) found only 9% decrease in diffuse activity with castration instead of the 98% found by Hoskins (**40**) for running activity.

In light of these considerable differences, it seems logical that different terms should be used to distinguish the two devices and the behaviors which they measure. Throughout this paper, the author has attempted to distinguish between running activity in the rotating drum, on the one hand, and diffuse activity in a cage or stabilimeter on the other As long as the terminology for these two distinct situations is the same, the notion will tend to persist that they are strictly comparable measures, which they are not.

Heredity and Age

Genetic Basis. Rundquist (**89**) by selective breeding has been able to get active and inactive strains of rats. The active strain is less easy to purify than the inactive strain. Selection for running produced strains in which there were measurable: (1) increases in number of successful matings, (2) increases in sizes of litters, and (3) decreases in the gestation period. The active rats also had a higher basal metabolic rate than the inactive strain. The selective breeding of these strains has been carried through 29 generations, with no change beyond the 12th (**9, 30**). Brody has concluded that the two strains differ with respect to a single gene which acts as a dominant in males and a recessive in females. This gene must act as an inhibitor, since none of the matings within the inactive strain produces active offspring, but on the other hand, active-strain matings produce individuals which vary from extreme inactivity to extreme activity. The genetic factors are somewhat obscured by environmental influences.

Age. Running activity of rats increases with age until the animals are about 80 days old, then is relatively constant until about 120 days, after which it gradually falls off till death (**72, 95, 101**). Richter (**73**) determined the amount of running, diffuse activity, and nest building as a function of age. The more active rats tend to have shorter lives than the inactive rats.

The Internal Environment

Nutrition

The daily running activity of the rat increases just prior to the normal time of feeding, even though the rat has been fed in what would otherwise be an inactive period (**72, 103**). This fact is probably explicable on the basis of hunger or metabolic changes connected with the daily 24-hour hunger rhythm initiated by constant feeding at a specific hour (**120**). Studies made by Richter (**73**) on generalized activity, however, show that the rat very probably has a two-hour hunger rhythm if he is allowed to have food constantly available.

If a rat is deprived of food for a period of time, its activity will tend to increase for as much as 96 hours. If deprived of food and water, it increases for 72 hours before it drops off (**73, 117**), probably due to weakness. Brobeck (**8**) showed that if food intake and environmental temperature are held constant, there is a negative correlation between activity and weight gain. Smith and Conger (**107**) varied the diet of rats by keeping the caloric value constant but changing the proportion

of fat or protein. They found that up to 56% of the caloric value of food may come from fat without reduction in spontaneous activity. Fifty per cent of animal protein, however, induces marked reduction in running. Following ingestion of protein there is a marked metabolic rise which is not present following ingestion of fats. This result extends Slonaker's (**105**) finding that a diet of 14 to 18% protein yields maximal spontaneous running. Hitchcock (**39**), however, fed rats eight grams of meat daily and observed little effect on running.

A protein-free diet for short periods of time is effective in increasing activity. But a low-protein diet over a long period of time depressed activity (**39**).

When allowed to select the amount of alcohol ingested, rats maintained their activity, although the forcing of alcohol by inclusion in the drinking water supply caused a reduction in activity (**35**).

Vitamins

Rats deprived of food, water, food and water, thiamin, and riboflavin showed increased activity until they were given ample amounts of the formerly scarce substance; then they were quiescent to a greater degree than normal until they had again 'caught up' (**117**). Limitation of Vitamin B is promptly followed by increased running activity in the rat; after five to ten days, however, activity is sharply decreased, long before any clinical signs of deficiency appear. This is accompanied by a drop in quantity of food ingested. When these rats are then given unlimited amounts of food and vitamins, activity decreases and remains low throughout the period of increased intake (**5**). These two papers do not completely corroborate Jackway's earlier finding (**47**) that slight deprivation of Vitamin B complex, not severe enough to cause any symptoms except retardation of growth and slight roughness of coat, results in a decrease in voluntary activity.

In a very suggestive paper Ziegler and Knudsen (**124**) showed that young white rats fed on a rachitic diet during infancy ran less after recovery from the rickets than normals. If, however, the rats were deprived of Vitamin D at an even earlier age by feeding the rachitic diet to their mothers during pregnancy, the surviving pups were more active after recovery than normals.

That deficiency is not always accompanied by increased activity is demonstrated by the fact that rats deprived of the magnesium ion show a consistent drop in activity without the initial rise usually accompanying deprivation (**117**). Smith and Smith (**108**) observed that rats fed a diet low in inorganic constituents gradually declined in vol-

untary running. The reduction in running occurred long before the appearance of impaired running ability.

Drugs

There has been interest in the effects of drugs on behavior as showing perhaps quantifiable action similar to that qualitatively produced in human beings. In general, analeptics stimulate activity (6), but have a depressant or diminishing effect with prolonged use. Thus cocaine, benzedrine, ephedrine, and propadrine increased running for single doses, but effects of tolerance showed when the specific drug was administered for several days (116, 123). Caffeine, metrazol, and picrotoxin depressed running activity, but enhanced restless activity as measured in diffuse activity cages (116). In a subsequent paper, Shulte and Tainter (91) have shown the differential temporal effects of the administration of caffeine, coramine, and metrazol on diffuse activity.

Running activity is increased by benzedrine sulphate and by Kola, but there is a decrement in activity just before feeding time 24 hours later (35, 92).

Phenobarbital administered intraperitoneally (69) to white rats depressed the nighttime running level, but had no effect on the daytime level which is normally much lower than that of the nighttime. These injected rats ate less, drank less, and gained less weight. Even after withdrawal of phenobarbital, the experimental rats were less active than the controls (49).

Other drugs used have been dinitrophenol and thiouracil. These are discussed below.

Endocrines

Adrenals. In a review of experiments dealing with endocrines up to 1927, Hoskins (40) reports that entire removal of the adrenals leads to marked reduction in running. This operation is very difficult to perform successfully, but in all cases in which there was little reduction in activity, subsequent autopsy showed hypertrophied adrenal particles. In 12 out of 27 adrenalectomized rats, Richter (79) was able to restore running partially by implants of adrenal gland to the ovaries. Subsequent removal of the grafts decreased activity again. Histological examination showed viable cortical tissue but no medulla. Increase of salt in the diet produced a definite but partial increase in activity. Ergographic studies show that while the absolute strength of muscles remains unaltered, the capacity for work following adrenalectomy is reduced to 8% of normal. Bilateral abdominal sympathectomy and

adrenal inactivation by section of the nervous connection reduced running activity slightly for a period of 10 days with no apparent effect after that time (1). Lacey (**54**) injected adrenalin into rats and found a decrease in total diffuse activity although restlessness as indicated by behavior in the normally inactive period increased markedly.

Thyroid. Hoskins (**40**) has reported no reduction in running resulting from thyroidectomy. This result has been questioned by Hall and Lindsay (**34**) who found 50% reduction using the same measure as that reported by Hoskins. Richter (**75**) has pointed out that the results obtained depend upon the amount of thyroid tissue remaining, for a very small piece is enough to sustain activity. Thyroidectomized rats may be re-activated by replacement therapy. However, if the quantity of thyroid administered is too great, the activity is once more reduced. That reduction in metaboiic rate as such is not responsible for decrease in activity was shown by increasing the metabolism of thyroidectomized and normal rats by dinitrophenol (**33**). Inhibition of the thyroid by thiouracil resulted in slower growth, lowered metabolism, increase in the length of interoestral periods, reduction of spontaneous activity and disruption of rhythmic patterns of activity (**61**). Rundquist's breed of active rats had a higher metabolic rate than his inactive strain (**90**).

Pituitary. Ablation of the pituitary leads to marked decrease in running activity. Injection of emulsion of fresh anterior lobe of the hypophysis reduced the decrease in activity, but only in one animal out of seven did this replacement therapy cause activity to remain at the original level. Viable hypophyseal transplants to the anterior chamber of the eye had only a slight stimulating effect (**81**).

When the stalk of the pituitary is severed in the female rat, the normal oestral periodicity is greatly slowed down. Five female rats showed oestral periods separated by the following numbers of days: 9.6, 11.2, 13.0, 14.3, 18.4. Richter explains this as being probably a whole multiple of the fundamental four-to-five-day ovulation cycle. This explanation is corroborated by the fact that ovariectomy eliminated all cycles (**77**).

The genetically hypopituitary animal, the dwarf rat, possesses a diurnal rhythm (**70**).

Liver. Normal function of the liver is important in maintaining activity. Ligation of the bile duct caused almost complete reduction in animals in which the duct remained occluded. Associated with this effect was widespread destruction of liver cells, considerable dilation of the bile duct, and other profound peritoneal effects (**80**). Partial hepatectomy was performed by Dugal and Ross (**17**) who found that the

25% of the liver remaining took the same length of time to regenerate as did activity to return to normal.

Pancreas. The decrease in activity following pancreatectomy is variable. Some rats show little loss, others great. This reduction is attributed to the inability to metabolize carbohydrates (**85**).

Ovaries. There are several measurable sex differences in daily running activity. The female goes through a cycle of activity of about four to five days' duration, the peak of which is reached at oestrus (**23, 118**). Hourly records of the female show variation in the distribution of the activity as a function of the time in the oestrus cycle (**103**). This cycle may be abolished by bilateral ovariectomy (**77, 87**), when the activity in the drum is decreased by as much as 98%, but it is restored by ovarian transplants (**87**). The presence of the uterus is not essential to this activity cycle, as is shown by the fact that hysterectomy has no permanent effect on running activity (**18, 19**). If copulation takes place during oestrus, the activity is reduced until pregnancy and lactation are over and the female is again in oestrus as shown by cornified epithelial cells or receptivity. Stimulation of the cervix during receptivity leads to pseudo-pregnancy with about a 10-day marked reduction in running, followed by gradually increasing activity until the previous level is reached in about six days (**104**).

The cycles of the ovariectomized rat may be restored in females or impressed in castrated males by several means, all of which involve supplanting the hormonal deficiency. Pregnancy urine of women seven to nine months pregnant, given in drinking water, was found to restore activity to near normal levels (**78**). Amniotin injected in spayed females and castrated males had a similar effect (**82**). The earlier discrepancy between results of other investigators, reported by Shirley, is possibly explicable by the findings of Young and Fish (**122**). They were able to restore activity to levels prevailing before gonadectomy by estrone administered in such a way that a constant source of the hormone was maintained. A certain small amount of estrone is necessary for the manifestation of running activity, but beyond that quantity, further estrone does not increase the activity nor alter its periodicity. These are similar to Heller's results (**36**) on male hormone. Castrated males show oestrus cyclic activity from ovarian transplants. Elderly males increase activity and lose weight following ingestion of an estrogenic fraction of chorionic gonadotropic extract (**41**).

Testes. The running activity of the male is generally much less than that of the female, and lacks its cyclical characteristics. Complete castration leads to marked reduction (about 98%) in running. Fractional

castration by Gans (27) of the following amounts of the testes of rats: 0, 1, 1½, 1¾, 2, led to the following decrements in running compared with his controls: 20%, 11%, 26%, 48%, 79%. Probably there was not time for regeneration to take place. Castration before the 12th day after birth leads to no activity reduction in the adult rat (26) but this statement has been questioned by Richter (76) who found the reduction in running almost as great as in rats castrated after sexual maturity. Richter ascribes this difference to the relative inactivity of Gans' controls. It is possible, however, that the findings of Hunt and Schlosberg (42, 43) would bear on the question. They used general diffuse activity cages and found little difference between the number of five-minute active periods of normal and castrated males. The running activity of both male and female castrated rats is increased by testicular grafts, provided these grafts are "functional takes" (87).

The relation between sexual drive and running activity is unclear. A male rat in a running drum near one female rat will show peaks of activity correlated with her oestral periods. If several females are in nearby drums, the total running of the male is increased but is no longer of the simple oestral periodicity (20, 106). On the other hand, the correlation between sex drive (as measured by copulatory tests and the Columbia obstruction box) and running activity is near zero (113). Rundquist (89) found that rats bred for activity were more fertile than those bred for inactivity. These facts are not contradictory but do require clarification.

External Environment

Light and Darkness

The rat displays more running and more diffuse activity in darkness than in light. This rhythm is of a 24-hour periodicity and is remarkably stable. The daily activity periods are determined by the internal rhythm of the animal, which rhythm is only gradually changed by the external conditions—light and temperature (10, 13, 14, 43, 48). Using a rhythm involving alternation of eight hours of dark and eight hours of light, Hemmingsen and Krarup were unable to abolish the 24-hour rhythm even in rats kept under these conditions from birth (37). By using six hours of light and warmth, alternating with six hours of cooler darkness, Browman (14) was able to impress a 12-hour rhythm on about 30 out of 32 rats. The rhythm was not stable, however, as shown by the fact that the longest period during which it persisted in any one rat was 37 days, in spite of the continual 12-hour cycle of light and heat.

The female rat shows a four-to-five-day oestrous cycle of activity

which can be somewhat altered by lighting in that constant light leads to longer intervals between peaks of oestrous activity and lowered daily activity (37). Browman (10) even found evidence of constant cornification (but not receptivity) in some of his animals. Subsequently he demonstrated that the intact visual apparatus was necessary for this shift to take place (11, 12). Females who were blinded showed the same 24-hour rhythm as those reared in constant darkness. Feeding schedules apparently had little effect on the running time. That the intensity of the constant light is important has been nicely demonstrated by Johnson (48) who found that the amount of change in the period of maximal activity of mice was related to the quantity of cage illumination. These results were obtained with mice in diffuse activity cages, since the mouse apparently exhibits even more dichotomous activity cycles than the rat. Working with male rats in diffuse activity cages, Hunt and Schlosberg (43) attempted to determine the source of the 24-hour activity rhythm; since castration did not abolish the periodicity they came to the conclusion that it was not based on the testes. Reversing the normal day and night periods caused the rats to shift their period of maximal activity in about a week.

In constant light, the female rat will show peaks of activity which are correlated with a daily cool period (13). This is also true of blinded rats. The pre-oestrus activity readings come at the end of the 12-hour cool period. It had previously been shown that in a hot room, rats will reduce their activity slightly (13, 97). Below about 40 degrees Fahrenheit the rat's activity is reduced also.

Temperature

The effect of temperature on the diffuse activity of two-day-old mice was determined by Stier (111). These young mice are still poikilothermic so that control of external temperature had direct effect on the body temperature. It was found that different straight-line Arrhenius plots would fit the several measures used: amount of activity, quiescence, and frequence of occurrence. Thus, he concluded that activity is a function of several factors.

Activity Following Confinement

The drive for activity hypothesized in several behavior theories has not been fully substantiated. Shirley (95) reports increased running activity after a rat has been confined one to two days, but after ten days of enforced activity, running decreases. Siegel (99) found an increase in the number of interruptions of a light beam following close confinement.

It is important to note that in neither case is the amount of activity under conditions of no confinement significantly different from the increase reported after confinement.

Skinner (100) has stated that "if any extensive activity is prohibited during part of a day, the remaining part shows a greater 'density' of activity per unit of time." Since he measured activity for a five to six hour period beginning at 3:00 a.m., it is possible that his conclusion is an artifact based upon the normal peak of the rat's running activity which occurs at about 2:00 a.m. After this time there is a gradual decrease in the amount run, a rhythm which will persist in constant darkness. Without further work substantiating the generality of the principle of greater activity following inactivity, it would be parsimonious to ascribe Skinner's results to the specific hours at which he recorded activity.

Miscellaneous Studies

Other studies using the running drum which might be mentioned are: the correlation between activity and errors in learning a maze is low (89, 96); while that between activity and time to traverse mazes is higher (68); rats given difficult discriminations run less than rats that are not made abnormal by difficult problems (60); a series of electroconvulsive shocks greatly reduced voluntary activity in rats (112).

Neural Control of Spontaneous Activity

The search for a neural center which controls activity has led to contradictory evidence. There is complete agreement that frontal lesions do increase activity and that bilateral lesions are more effective than uni-lateral lesions. The hyperactivity usually takes two to three weeks to emerge. The animals on which most of the work has been done are rats, cats, and monkeys.

The first observation of the effect of cortical lesions on activity has been attributed to several clinicians and experimenters. In 1920 Lashley (57) quantified the change in activity of the rat by using a running drum and noted that only frontoparietal lesions both increased the number of hours of running and decreased the time spent in resting. Jacobsen (45, 46) noted increased restlessness and general activity following frontal destruction in the monkey, and Langworthy and Kolb (56) described the behavior of cats with heightened restlessness. Since 1937, considerable evidence has been gathered concerning the effect on spontaneous activity produced by lesions in the brain.

Rats

In the rat, the unilateral removal of the frontal pole did not appear to augment running activity. The inactive rats became hyperactive while the active rats did not increase their running relatively so much. Bilateral ablation of the frontal poles was much more effective in increasing running than the unilateral (3, 83). Beach (4) measured running for 30 days before and 50 days after electrolytic destruction of varying amounts of the corpus striatum. Activity increased postoperatively in one animal, decreased in two, and was unchanged in two others. "No relationship between the magnitude of lesion and effects on activity could be determined. In the rat, the striatum evidently does not exert a controlling effect upon running activity as measured in this experiment." On the other hand, Richter and Hines (84) found that monkeys with unilateral striatal lesions immediately had greatly increased activity, and Mettler (64) reports hyperactivity in cats following striatal lesions.

According to Mettler (63, 64) when the striatum is injured, hyperkinesia is the rule. He asserts that the striatum is an inhibitory mechanism; " . . . stimulation of it produces inhibition and removal of it engenders evidence of motor release. It stands on the one hand between the cortex and the final common path as part of the route through which the cortex may exert an inhibitory effect and, on the other hand, it operates between the thalamus and lower motor mechanisms in the automatic inhibition incident to 'unconscious activity'." If the cerebral cortex is totally removed, the decorticated animal does not exhibit incessant activity but shows an inability to initiate or inhibit movement suddenly (65).

Cats

Cats with bilateral one-stage removal of the rostral portions of the cerebral hemispheres were noted by Magoun and Ranson (59) to be almost continually walking about. Langworthy and Richter (55) recorded increase in activity from 27 to 61 units from unilateral operation and to 399 units for the bilateral removal of motor cortex, premotor cortex, and possibly a small tip of the corpus striatum in cats.

Monkeys

In monkeys, Richter and Hines (84) found that bilateral removal of areas 8, 10, 11, and 12 had little effect on activity, while that of 9 did (62). On the other hand, Kennard and Ectors (50) reported increased activity following removal of area 8 alone. These results are not con-

tradictory in view of the method of measurement of activity. Richter and Hines (84) attached the monkey by a chain to a short steel rod projecting from an axle. Movement of the monkey caused the rod to advance a counter. The method of recording activity used by Kennard *et al.* was a pneumatically mounted diffuse activity cage.

Generally the activity builds up in the course of two to three weeks following an operation. Ruch and Shenkin (88), however, report that lesions in area 13 (of Walker) consistently produce hyperactivity within the second post-operative day. Richter and Hines (84) also report such immediate hyperkinesia when the monkey has striatal lesions.

Kennard *et al.* (51) have stressed the visual role in hyperactivity in monkeys. "Hyperactivity is markedly affected by visual stimuli. It disappears in the dark or when the animals have been deprived of vision either by enucleation of the eyes of by bilateral lobectomy. Absence of auditory stimuli has not the same effect."

A decrease in activity was noted by Barris (2) following "bilateral one-stage removal of the rostral portions of the neo-cortex of cats." Kennard *et al.* reported that "hypermotility in monkeys and chimpanzees is related to lesions of the rostral portions of area 6 and to area 8" (51).

Summary

The literature of the last twenty years concerning activity in animals has been reviewed. The methods, results and concepts of activity have been summarized and appraised.

Several methods have been in use:

1. The running drum: this apparatus yields high reliability for measures taken on a particular drum, but there are also inconsistencies from one drum to another and from one experimenter's design of drum to another.
2. The diffuse activity cage: a cage mounted on tambours or springs. The record which it gives varies widely from one cage to another.
3. Several miscellaneous mechanical and observational methods, which have not been extensively used.

There are considerable individual differences in running activity as well as variability of one animal's activity from time to time. The individual differences are due in part to heredity, but are somewhat complicated by environmental influences. Intra-animal variability can be ascribed to several factors: running increases during hunger and most deprivations, during darkness and cool periods, and during oestrus. In various kinds of endocrine imbalance or deficiency, there is usually a decrease in activity—a marked decrease in the running drum but only a small decrement in the diffuse activity cage.

Running activity and diffuse activity are sometimes affected in the same way, sometimes differentially. Both reach a maximum during the cool or dark part of a 24-hour cycle. Some of the analeptic drugs stimulate both kinds of activity, but other drugs may increase diffuse activity while decreasing running activity.

Injury to the brain affects activity. In particular, lesions of the frontal cortex heighten activity, and bilateral lesions cause a greater increase than unilateral injury. Still, it is not yet clear whether there is a specific activity center in the brain and, if so, where it is.

There is now a very large body of data concerning animal activity, but it needs further definition and interpretation. Particularly needed is a clarification of the concept of activity in relation to the method of measuring it. Most treatments of the subject tend to regard activity as a single entity. Yet, in some cases, where comparable measures of activity are available from different devices, running drum and diffuse activity cage, the results are not the same. Activity, it would then appear, does not constitute a single behavior category which can be measured with any instrument but must be considered, for the present at least, in terms of its method of measurement.

BIBLIOGRAPHY

1. Bacq, Z. M. The effects of abdominal sympathectomy, adrenal inactivation and removal of the stellate ganglia on the spontaneous activity of the albino rat. *Endocrinology*, 1931, **15**, 34–40.
2. Barris, R. W. Cataleptic symptoms following bilateral cortical lesions in cats. *Amer. J. Physiol.*, 1937, **119**, 213–220.
3. Beach, F. A. Effects of brain lesions upon running activity in the male rat. *J. comp. Psychol.*, 1941, **31**, 145–179.
4. Beach, F. A. Effects of lesions to corpus striatum upon spontaneous activity in the male rat. *J. Neurophysiol.*, 1941, **4**, 191–195.
5. Bloomfield, A., & Tainter, M. L. The effect of vitamin B deprivation on spontaneous activity of the rat. *J. Lab. clin. Med.*, 1943, **28**, 1680–1690.
6. Boughton, L. L. The effect of life cycle therapeutic dosage administration of drugs to albino rats. II. On activity, maze learning and relearning. *J. Amer. pharm. Ass.*, 1942, **31**, 240–244.
7. Bousfield, W. A., & Mote, F. A. The construction of a tilting activity cage. *J. exp. Psychol.*, 1943, **32**, 450–451.
8. Brobeck, J. R. Effects of variations in activity, food intake and environmental temperature on weight gain in the albino rat. *Amer. J. Physiol.*, 1945, **143**, 1–5.
9. Brody, E. G. The genetic basis of spontaneous activity in the albino rat. *Comp. Psychol. Monogr.*, 1942, **17**, No. 5. Pp. 24.
10. Browman, L. G. Light in its relation to activity and estrous rhythms in the albino rat. *J. exp. Zool.*, 1937, **75**, 375–388.
11. Browman, L. G. The effect of bilateral optic enucleation on the voluntary muscular activity of the albino rat. *J. exp. Zool.*, 1942, **91**, 331–344.
12. Browman, L. G. The effect of bilat-

eral optic enucleation upon the activity rhythms of the albino rat. *J. comp. Psychol.*, 1943, **36**, 33–46.

13. BROWMAN, L. G. The effect of controlled temperatures upon the spontaneous activity rhythms of the albino rat. *J. exp. Zool.*, 1943, **94**, 477–489.
14. BROWMAN, L. G. Modified spontaneous activity rhythms in rats. *Amer. J. Physiol.*, 1944, **142**, 633–637.
15. CURTIS, Q. F. Diurnal variation in the free activity of sheep and pig. *Proc. Soc. exp. Biol. & Med.*, 1937, **35**, 566–567.
16. DORCUS, R. M. A new device for studying motivation in rats. *J. comp. Psychol.*, 1934, **18**, 149–151.
17. DUGAL, L. P., & ROSS, S. Effet de l'ablation partielle du foie sur l'activité spontanée du rat blanc. *Rev. Canad. Biol.*, 1943, **2**, 435–441.
18. DURRANT, E. P. Studies on vigor. XI. Relation of hysterectomy to voluntary activity in the white rat. *Amer. J. Physiol.*, 1927, **82**, 14–18.
19. DURRANT, E. P. Relation of hysterectomy of long standing to voluntary activity in the white rat. *Amer. J. Physiol.*, 1931, **97**, 519. (Abstr.)
20. DURRANT, E. P. Influence of the female white rat in bodily activity of the male. *Amer. J. Physiol.*, 1935, **113**, 37. (Abstr.)
21. FARRIS, E. J., & ENGUALL, G. Turntable for exercising rats. *Science*, 1939, **90**, 144.
22. FARRIS, E. J. Apparatus for recording cyclical activity in the rat. *Anat. Rec.*, 1941, **81**, 357–361.
23. FARRIS, E. J. Leucopenia associated with normal estrous in the albino rat. *Anat. Rec.*, 1942, **82**, 147–151.
24. FARRIS, E. J. Breeding of the rat. Ch. 1 in J. Q. Griffith, Jr. and E. J. Farris (Eds.), *The rat in laboratory investigation*. Philadelphia: Lippincott, 1942. Pp. 1–17.
25. FREDERICSON, E. The theory of psychomotion as applied to a study of temperament. *J. comp. Psychol.*, 1946, **39**, 77–89.
26. GANS, H. M. Studies in vigor. XIII. The effect of early castration on the voluntary activity of male albino rats. *Endocrinology*, 1927, **11**, 141–144.
27. GANS, H. M. Studies on vigor. Effect of fractional castration on the voluntary activity of male albino rats. *Endocrinology*, 1927, **11**, 145–148.
28. GEIER, F. M. The measurement of tension in the rat. A contribution to method. *J. comp. Psychol.*, 1942, **34**, 43–49.
29. GEIER, F. M., & TOLMAN, E. C. Goal distance and restless activity. I. The goal gradient of restless activity. *J. comp. Psychol.*, 1943, **35**, 197–204.
30. GRAVES, E. A. The genetic basis of activity in the albino rat. *Psychol. Bull.*, 1937, **34**, 757–758. (Abstr.)
31. GRAY, W. L. *The effects of forced activity on maze learning and the selection and consumption of food by rats*. Unpublished Ph.D. Thesis, Johns Hopkins Univ., 1933.
32. HALL, C. S. Emotional behavior in the rat. III. The relationship between emotionality and ambulatory activity. *J. comp. Psychol.*, 1936, **22**, 345–352.
33. HALL, V. E., & LINDSAY, M. The effect of dinitrophenol on the spontaneous activity of the rat. *J. Pharm. exp. Therap.*, 1934, **51**, 430–434.
34. HALL, V. E., & LINDSAY, M. The relation of the thyroid gland to the spontaneous activity of the rat. *Endocrinology*, 1938, **22**, 66–72.
35. HAUSMANN, M. F. The behavior of albino rats in choosing food and stimulants. *J. comp. Psychol.*, 1932, **13**, 279–309.
36. HELLER, R. E. Spontaneous activity in male rats in relation to testis

hormone. *Endocrinology*, 1932, **16**, 626–632.

37. HEMMINGSEN, A. M., & KRARUP, N. B. Rhythmic diurnal variations in the oestrus phenomena of the rat and their susceptibility to light and dark. Det. Kgl. Danske Videnskabernes Selskab., *Biologiske Meddelelser*, 1937, **13**, 1–61.
38. HERRING, V. V., & BRODY, S. Growth and development. XLIII. Diurnal metabolic and activity rhythms. *Univ. Missouri Agri. Exp. Station Res. Bull.* 274, 1938 (not seen).
39. HITCHCOCK, F. A. The effect of low protein and protein-free diets and starvation on the voluntary activity of the albino rat. *Amer. J. Physiol.*, 1928, **84**, 410–416.
40. HOSKINS, R. G. Studies on vigor. XVI. Endocrine factors in vigor. *Endocrinology*, 1927, **11**, 97–105.
41. HOSKINS, R. G., & BEVIN, S. The effect of fractionated chorionic gonadotropic extract on spontaneous activity and weight of elderly male rats. *Endocrinology*, 1941, **27**, 927–931.
42. HUNT, J. McV., & SCHLOSBERG, H. General activity in the male white rat. *J. comp. Psychol.*, 1939, **28**, 23–38.
43. HUNT, J. McV., & SCHLOSBERG, H. The influence of illumination upon general activity in normal, blinded and castrated male white rats. *J. comp. Psychol.*, 1939, **28**, 285–298.
44. IRWIN, O. C. Effect of strong light on the body activity of newborns. *J. comp. Psychol.*, 1941, **32**, 233–236.
45. JACOBSEN, C. F. A study of cerebral function in learning. The frontal lobes. *J. comp. Neurol.*, 1931, **52**, 271–340.
46. JACOBSEN, C. F. Studies of cerebral function in primates. *Comp. Psychol. Monogr.*, 1936, **13**, No. 3. Pp. 68.
47. JACKWAY, I. Voluntary activity in the rat as related to the intake of whole yeast. *J. comp. Psychol.*, 1938, **26**, 157–162.
48. JOHNSON, M. S. Effect of continuous light on periodic spontaneous activity of white-footed mice. *J. exp. Zool.*, 1939, **82**, 315–328.
49. JONES, M. R. The effect of phenobarbital on food and water intake, activity level, and weight gain in the white rat. *J. comp. Psychol.*, 1943, **35**, 1–10.
50. KENNARD, MARGARET A., & ECTORS, L. Forced circling in monkey following lesions of the frontal lobes. *J. Neurophysiol.*, 1938, **1**, 45–54.
51. KENNARD, MARGARET A., SPENSER, SUSAN, & FOUNTAIN, G., JR. Hyperactivity in monkeys following lesions of the frontal lobes. *J. Neurophysiol.*, 1941, **4**, 512–524.
52. KREEZER, G. L. Technics for the investigation of psychological phenomena in the rat. Ch. 10 in J. Q. Griffith, Jr. and E. J. Farris (Eds.), *The rat in laboratory investigation*. Philadelphia: Lippincott, 1942. Pp. 199–273.
53. LACEY, O. L. A revised procedure for the calibration of the activity wheel. *Amer. J. Psychol.*, 1944, **57**, 412–420.
54. LACEY, O. L. The dependence of behavior disorders in the rat upon blood composition. *J. comp. Psychol.*, 1945, **38**, 277–284.
55. LANGWORTHY, O. R., & RICHTER, C. P. Increases in spontaneous activity aroused by frontal lobe lesions in cats. *Amer. J. Physiol.*, 1939, **126**, 158–161.
56. LANGWORTHY, O. R., & KOLB, L. C. The experimental production in the cat of a condition simulating pseudo-bulbar palsy. *Amer. J. Physiol.*, 1935, **111**, 571–577.
57. LASHLEY, K. S. Studies of cerebral function in learning. II. *Psychobiology*, 1920, **2**, 55–136.
58. LEVY, D. M. On the problems of

movement restraint, tics, stereotyped movements, hyperactivity. *Amer. J. Orthopsychiat.*, 1944, **14**, 644–671.

59. MAGOUN, H. W., & RANSON, S. W. The behavior of cats following bilateral removal of the rostral portion of the cerebral hemispheres. *J. Neurophysiol.*, 1938, **1**, 39–44.
60. MAIER, N. R. F., & WAPNER, S. Studies of abnormal behavior in the rat. *J. comp. Psychol.*, 1944, **37**, 151–158.
61. MANN, C. W. The effect of thiouracil upon the heart rate, estrous cycle and spontaneous activity of the white rat. *J. Psychol.*, 1945, **20**, 91–99.
62. MASSIMY, R., & CHEVALLIER, R. J. Les effets, chez le singe, de l'ablation préfrontal unilatérale; modifications de l'activité, du mode réactionnel et des réflexes. *C. R. Soc. Biol. Paris*, 1942, **136**, 103–106.
63. METTLER, F. A. Relation between pyramidal and extra-pyramidal function. *Res. Publ. Ass. nerv. ment. Dis.*, 1942, **21**, 150–227.
64. METTLER, F. A., & METTLER, C. C. The effects of striatal injury. *Brain*, 1942, **65**, 242–255.
65. METTLER, F. A., METTLER, C. C., & CULLER, E. A. Effects of total removal of cerebral cortex. *Arch. Neurol. Psychiat., Chicago*, 1935, **34**, 1238–1249.
66. MORGAN, C. T. *Physiological psychology.* New York: McGraw-Hill, 1943.
67. MUNN, N. L. *An introduction to animal psychology: the behavior of the rat.* New York: Houghton Mifflin, 1933. Pp. 50–78.
68. OMWAKE, L. The activity and learning of white rats. *J. comp. Psychol.*, 1933, **16**, 275–285.
69. OMWAKE, L. The influence of barbital on the activity and learning of white rats. *J. comp. Psychol.*, 1933, **16**, 317–325.
70. OSBORN, C. M. Spontaneous diurnal activity in a genetically hypopituitary animal, the dwarf rat. *Anat. Rec.*, 1940, **78**, Suppl. p. 137.
71. PARK, O., & WOODS, L. P. A modified Hemmingsen-Krarup mammalian activity recorder. *Proc. Soc. exp. Biol. Med.*, 1940, **43**, 366–370.
72. RICHTER, C. P. A behavioristic study of the activity of the rat. *Comp. Psychol. Monogr.*, 1922, **1**, 1–55.
73. RICHTER, C. P. Animal behavior and internal drives. *Quart. Rev. Biol.*, 1927, **2**, 307–343.
74. RICHTER, C. P. Symposium: Contributions of psychology to the understanding of problems of personality and behavior. IV. Biological foundations of personality differences. *Amer. J. Orthopsychiat.*, 1932, **2**, 345–354.
75. RICHTER, C. P. The role played by the thyroid gland in the production of gross body activity. *Endocrinology*, 1933, **17**, 73–87.
76. RICHTER, C. P. The effect of early gonadectomy on the gross body activity of rats. *Endocrinology*, 1933, **17**, 445–450.
77. RICHTER, C. P. Cyclical phenomena produced in rats by section of the pituitary stalk and their possible relation to pseudo-pregnancy. *Amer. J. Physiol.*, 1933, **106**, 80–89.
78. RICHTER, C. P. Pregnancy urine given by mouth to gonadectomized rats: its effect on spontaneous activity and on the reproductive tract. *Amer. J. Physiol.*, 1934, **110**, 499–512.
79. RICHTER, C. P. The spontaneous activity of adrenalectomized rats treated with replacement and other therapy. *Endocrinology*, 1936, **20**, 657–666.
80. RICHTER, C. P., & BENJAMIN, J. A., JR. Ligation of the common bile duct in the rat. *Arch. Path.*, 1934, **18**, 817–826.

81. RICHTER, C. P., & ECKERT, J. F. The effect of hypophyseal injection and implants on the activity of hypophysectomized rats. *Endocrinology*, 1937, **21**, 481–488.
82. RICHTER, C. P., & HARTMAN, C. G. The effect of injection of amiotin on the spontaneous activity of gonadectomized rats. *Amer. J. Physiol.*, 1934, **108**, 136–143.
83. RICHTER, C. P., & HAWKES, C. D. Increased spontaneous activity and food intake produced in rats by removal of frontal poles of the brain. *J. Neurol. Psychiat., Chicago*, 1939, **2**, 231–242.
84. RICHTER, C. P., & HINES, M. Increased spontaneous activity produced in monkeys by brain lesions. *Brain*, 1938, **61**, 1–16.
85. RICHTER, C. P., & SCHMIDT, E. C. H., JR. Behavior and anatomical changes produced in rats by pancreatectomy. *Endocrinology*, 1939, **25**, 698–706.
86. RICHTER, C. P., & WANG, G. H. New apparatus for measuring the spontaneous motility of animals. *J. Lab. Clin. Med.*, 1926, **12**, 289–292.
87. RICHTER, C. P., & WISLOCKI, G. B. Activity studies on castrated male and female rats with testicular grafts, in correlation with histological studies of the grafts. *Amer. J. Physiol.*, 1928, **76**, 651–660.
88. RUCH, T. C., & SHENKIN, H. A. The relation of area 13 on the orbital surface of the frontal lobe to hyperactivity and hyperphagia in monkeys. *J. Neurophysiol.*, 1943, **6**, 349–360.
89. RUNDQUIST, E. A. Inheritance of spontaneous activity in rats. *J. comp. Psychol.*, 1933, **16**, 415–438.
90. RUNDQUIST, E. A., & BELLIS, C. J. Respiratory metabolism of active and inactive rats. *Amer. J. Physiol.*, 1933, **106**, 670–675.
91. SCHULTE, J. W., TAINTER, M. L., & DILLE, J. M. Comparison of different types of central stimulation from analeptics. *Proc. Soc. exper. Biol. Med.*, 1939, **42**, 242–248.
92. SEARLE, L. V., & BROWN, C. W. Effect of subcutaneous injections of benzedrine sulphate on the activity of white rats. *J. exper. Psychol.*, 1938, **22**, 480–490.
93. SEARLE, L. V., & BROWN, C. W. Effect of variation in the dose of benzedrine sulphate on the activity of white rats. *J. exp. Psychol.*, 1938, **22**, 555–563.
94. SHIRLEY, MARY. Studies in activity. I. The consistency of the revolving drum method of measuring the activity of the rat. *J. comp. Psychol.*, 1928, **8**, 23–38.
95. SHIRLEY, MARY. Studies in activity. II. Activity rhythms, age and activity, activity after rest. *J. comp. Psychol.*, 1928, **8**, 159–186.
96. SHIRLEY, MARY. Studies in activity. IV. The relation of activity to maze learning and brain weight. *J. comp. Psychol.*, 1928, **8**, 187–195.
97. SHIRLEY, MARY. Spontaneous activity. *Psychol. Bull.*, 1929, **26**, 341–365.
98. SIEGEL, P. S. A simple electronic device for the measurement of the gross bodily activity of small animals. *J. Psychol.*, 1946, **21**, 227–236.
99. SIEGEL, P. S. Activity' level as a function of physically enforced inaction. *J. Psychol.*, 1946, **21**, 285–291.
100. SKINNER, B. F. The measurement of "spontaneous activity." *J. gen. Psychol.*, 1933, **9**, 3–24.
101. SLONAKER, J. R. The normal activity of the white rat of different ages. *J. comp. Neurol. Psychol.*, 1907, **17**, 342–359.
102. SLONAKER, J. R. Description of apparatus for recording the activities of small mammals. *Anat. Rec.*, 1908, **2**, 116–122.
103. SLONAKER, J. R. Analysis of daily

activity of the albino rat. *Amer. J. Physiol.*, 1925, **73**, 485–503.

104. SLONAKER, J. R. Pseudopregnancy in the albino rat. *Amer. J. Physiol.*, 1929, **89**, 406–416.
105. SLONAKER, J. R. Effect of different per cents of protein in the diet. II. Spontaneous activity. *Amer. J. Physiol.*, 1931, **96**, 557–561.
106. SLONAKER, J. R. Sex drive in rats. *Amer. J. Physiol.*, 1935, **112**, 176–181.
107. SMITH, E. A., & CONGER, R. M. Spontaneous activity in relation to diet in the albino rat. *Amer. J. Physiol.*, 1944, **142**, 663–665.
108. SMITH, P. K., & SMITH, A. H. The effect of a diet low in inorganic constituents on the voluntary activity of the albino rat. *Abstr. Proc. Amer. physiol. Soc.*, 1934.
109. SMITH, K. U. An accurate method of recording activity in animals. *J. gen. Psychol.*, 1942, **27**, 355–358.
110. STEWART, C. C. Variations in daily activity produced by alcohol and by changes in barometric pressure and diet, with a description of recording methods. *Amer. J. Physiol.*, 1898, **1**, 40–56.
111. STIER, T. J. B. "Spontaneous activity" of mice. *J. gen. Psychol.*, 1930, **4**, 67–101.
112. STONE, C. P. Effects of electro-convulsive shocks on daily activity of albino rats in revolving drums. *Proc. Soc. exp. Biol., N. Y.*, 1946, **61**, 150–151.
113. STONE, C. P., & BARKER, R. B. Spontaneous activity, direct and indirect measures of sexual drives in adult male rats. *Proc. Soc. exp. Biol., N. Y.*, 1934, **32**, 195–199.
114. SZYMANSKI, J. S. Eine methode zur Einversuchung der Ruhe und Aktivitätsperioden bei Tieren. *Arch. f. d. ges. Physiol.*, 1914, **158**, 343–385.
115. SYZMANSKI, J. S. Activität und Ruhe bei Tieren und Menschen. *Z. allg. Physiol.*, 1920, **18**, 105–162.
116. TAINTER, M. L. The effects of certain analeptic drugs on spontaneous running activity of the white rat. *J. comp. Psychol.*, 194?, **36**, 143–155.
117. WALD, G., & JACKSON, B. Activity and nutritional deprivation. *Proc. nat. Acad. Sci., Wash.*, 1944, **30**, 255–263.
118. WANG, G. H. The relation between spontaneous activity and estrous cycle in the white rat. *Comp. Psychol. Monogr.*, 1923, **2**, 1–27.
119. WELSH, J. H. Diurnal rhythms. *Quart. Rev. Biol.*, 1938, **13**, 123–139.
120. WERTHESSEN, N. T. The significance of sub-normal respiratory quotient values induced by controlled feeding of the rat. *Amer. J. Physiol.*, 1937, **120**, 458–465.
121. WILBUR, K. M. A method for the measurement of activity of small animals. *Science*, 1936, **84**, 274.
122. YOUNG, W. C., & FISH, W. R. The ovarian hormones and spontaneous running activity in the female rat. *Endocrinology*, 1945, **36**, 181–189
123. ZIEVE, L. Effects of benzedrine on activity. *Psychol. Rec.*, 1937, **1**, 393–396.
124. ZIEGLER, L. H., & KNUDSON, A. Activity after recovery from rickets. *J. comp. Psychol.*, 1936, **22**, 199–217.

SAMPLING IN THE REVISION OF THE STANFORD-BINET SCALE

ELI S. MARKS
National Office of Vital Statistics

In another paper (4) the writer attempts to point out the biases which may arise through types of sampling procedure quite common in psychological research. The present analysis is devoted to another effect of sampling methods commonly used in psychology—namely, the substantial increase in sampling error which results when "cluster" methods of sampling are used. It should be noted that this is not a criticism of the cluster type of sampling. Cluster sampling is an extremely valuable device and makes feasible many studies which would otherwise be completely impossible. However, the use of cluster techniques implies substantial modifications in our formulae for sampling error and psychologists are, in general, not familiar with these modifications. Unfortunately, much of the important work in the field appears in sources which are relatively inaccessible to psychologists. Ignoring the effects of cluster sampling on measures of sampling error has undoubtedly resulted in attaching importance to results which are statistically insignificant. In the testing field, failure to allow for cluster sampling has probably caused us to attach a measure of precision to our results considerably in excess of that warranted by sound statistical techniques.

Cluster sampling almost always involves an increase in sampling error as compared with unrestricted random sampling of the same number of cases. It is, of course, possible to obtain a lower sampling error with cluster sampling than with unrestricted random sampling if we make up our clusters for this purpose. However, the main reason for the use of cluster sampling is to permit the sampling of previously existing groups (the clusters) and, in most cases, the use of a previously existing grouping of the population involves a positive intraclass correlation of the variable studied, i.e., our existing groups are almost always more homogeneous internally than groups of the same size made up by random selection of individuals from the population. It is the existence of positive intraclass correlation which cuts down the amount of independent information available from a cluster sample of a specified size and occasions the substantial increase in sampling error usually associated with this sampling method. The present analysis is designed to emphasize the substantial increase in sampling error which results from relatively small intraclass correlations. While this phenomenon is quite familiar to sampling statisticians, psychologists are rather generally

unaware of the marked disturbances of sampling error calculations and tests of significance introduced by the use of cluster sampling when a positive intraclass correlation exists.

Although methods resembling cluster sampling are quite common in psychological research, very few psychological studies have used sampling designs which permit us to determine the standard error of the mean or of other sample statistics. As a matter of fact, it is difficult to find a study where analysis of the sampling error formulae used is not complicated by the presence of a non-measurable design (one in which the sampling probabilities are unknown). Some of the difficulties in the use of non-measurable designs are explored in a paper by McNemar (**2**) which discusses accidental sampling and purposive sampling as well as such measurable designs as unrestricted random sampling and stratified sampling.

The Sampling Plan of the Stanford-Binet

The writer has, therefore, not attempted to find a study with a measurable design, but has selected for analysis the sample used in the revision of the Stanford-Binet. This sample has been selected for analysis principally because the widespread use of the revised Stanford-Binet makes the problems involved in its standardization extremely important in spite of the lapse of a decade since the revision was completed. The revision of the Stanford-Binet is also a good example for our purposes because (a) it was an extensive project, involving a relatively large number of subjects and the expenditure of considerable amounts of time, effort and money and (b) the purposes of the sample were explicitly formulated and clearly stated by the authors of the revised Stanford-Binet.

The reader should bear in mind that the present analysis is on an "as if" basis. The Stanford-Binet sampling design does not yield statistics with measurable standard errors and no amount of statistical manipulation can overcome this defect. The cure lies not in statistical formulae but in more careful sampling techniques in future investigations. However, the use of measurable sampling designs in psychological research will almost inevitably mean cluster sampling of some sort since any other approach will be beyond the limited resources usually available to psychologists. Thus, an examination of cluster sampling, even on an "as if" basis, is extremely pertinent to the future of any psychological research which involves statistical techniques.

In my analysis I have relied entirely upon statements and data published in Terman and Merrill (**6**) and McNemar (**3**). Since the data

required for this analysis have not been published in full detail, I have been forced to use approximations at several points. Inquiry indicated that more detailed data could not be furnished without considerable expenditure of time and effort. Since the approximations used in this paper are satisfactory for purposes of illustration and since the sampling techniques used in the revision of the Stanford-Binet preclude a completely accurate determination of error even if the detailed data were available, this deficiency is not serious. In nearly every case, the effect of the approximation used has been to understate the sampling error.

In revising the Stanford-Binet, the major objective was to construct scales "so standardized for difficulty as to yield mean I.Q.'s of approximately 100 at all age levels" (Terman, in **3**, p. 3). The authors of the revision realized that their success in this objective was dependent upon securing a measure of the distribution of test scores in the general population (or in a satisfactory sample of the population). The sample was restricted to "American born" subjects of the "white race" in the age range from 1½ years to 18 years. Terman notes that "elaborate precautions were taken to make the sampling as representative of the entire population as circumstances permitted" (**3**, p. 6).

According to Terman and Merrill this was done by selecting "17 different communities in 11 states" (**6**, p. 12). They note that: "The selection of localities for the second year's testing was based upon certain considerations in regard to sampling which had resulted from a study of the socio-economic level of the first 1500 subjects." These considerations were what the authors viewed as an inadequate representation of the rural group and a difference between the occupational distribution of fathers of the cases tested and the occupational distribution of all employed U. S. males. In the second year's testing, therefore, the authors of the Stanford-Binet revision "took care to include several additional rural communities" (**6**, p. 14) Neither McNemar nor Terman and Merrill give further details on the method of selecting the communities. It seems evident that selection was not on the basis of random sampling (neither simple random sampling nor random sampling within strata). As a matter of fact the term "community" is not defined clearly enough to permit a rigorous statement of the primary sampling units used. Nevertheless, we can visualize our population as being composed of "communities" (undefined but definable), so that the entire population of the United States can be broken up into a fairly large number (probably over 3000) of communities.

Within each community different procedures were followed for cases in the three age groups—1½ to 5½ inclusive; 6 to 14 inclusive; and 15 to 18 inclusive. These groups were sampled as follows:

1. *The group aged 6 to 14.* Schools of "average social status" were selected in each community (method of selecting schools not further specified) and within each school all of the children between the ages 6 to 14 who were within one month of a birthday were taken, regardless of grade placement (6, p. 15). This sampling procedure is, then, a subsampling of subclusters with a 100 percent sample take within the subcluster.

2. *The group aged 15 to 18.* Subjects were selected so that "the advanced group would be as nearly as possible continuous with the intermediate, with no break between fourteen and fifteen years. The compulsory school age was taken into account, the general character of the population, and the type of secondary education that was offered. In each community the school census was consulted to determine the amount of elimination after age fourteen. We made certain that some of the twelve-, thirteen-, and fourteen-year-olds who had gone to high school were included, also some of the slow fifteen- and sixteen-year-olds who were still in intermediate school. A few cases who had graduated from high school were included and a few who had dropped out of school without completing high school. These out-of-school groups were sampled by choosing siblings of school children in numbers proportional to the amount of elimination at ages above fourteen." (Sampling in this group was actually a rough type of "quota" sampling.)

3. *The group aged 1½ to 5½.* This group was sampled in much the same manner as the out-of-school cases in the group aged 15 to 18. The authors "chose as far as possible younger sibs of the school groups." Children were secured by use of birth records, school census, school siblings, kindergartens, well baby clinics, day nurseries, nursery schools and "personal report." Use of the various sources differed from community to community. "Great care was exercised in the large population centers to include representative groups; if a school in a suburban district which had been chosen as average on the advice of superintendent and counselors seemed to include too large percentage of higher occupational groups it was offset by a tenement district center." The authors state that "in the smaller communities, from seventy-five to eighty percent of the pre-school child population of appropriate age was examined" (6). In the published tabulations results for children aged 1½ were omitted and further references deal only with the sample of children two years of age or over.

It may be noted that the population sampled is limited to individuals within one month of a birthday (or half-year birthday for children under six). The population is also limited to American-born white persons and, in the age range six to 14, to children attending school. These limitations do not affect the propriety of generalization from a sample to the population so defined. The limitations may affect generalization from the sample to all native-born white persons aged two to 18. This is not, however, of primary concern in this paper. Limitations on generalization resulting from the use of sub-populations are discussed by the writer in another article (4). For our present purposes, it is sufficient to accept the population, as defined.

To summarize, the sampling plan of the Stanford-Binet revision

involved: (a) sampling of "communities" from the aggregate of all United States communities; (b) the subsampling of schools for children aged six to 14 and taking all children (in the population as defined above) in the selected schools; (c) the subsampling of other members of the defined population from the "community" without any intermediate subsampling of schools but with the use of a rough type of "quota" sampling.

Biases and Variance in the Stanford-Binet Sampling

The above is only an approximate statement since it is extremely hard to formulate exactly the sampling plan used. The method of selection at each stage of sampling has not been specified above. It seems likely, however, that the sampling error of the plan used is greater than the error which would be involved in random sampling of "communities" with equal probability of selection and no subsampling (i.e., a plan which would take all persons in the communities selected).

It is obvious that a sampling plan not involving subsampling will have a lower sampling error for the same number of clusters sampled than a plan which did involve subsampling. The assumption that the community sampling actually used involved a larger sampling error than random selection is not as clear cut. Actually the sampling resembled "purposive sampling" or "quota sampling" but it does not appear to conform even to the rather loose requirements of these two techniques.

In discussing purposive sampling Neyman (5) developed certain hypotheses which, if satisfied, would make the estimate commonly used in this method the "best linear estimate" (i.e., an unbiased linear estimate with variance less than that of any other linear estimate). Neyman notes that:

> If these hypotheses are not satisfied, which I think is a rather general case, we are not able to appreciate the accuracy of the results obtained. Thus this is not what I should call a representative method. Of course it may sometimes give perfect results, but these will be due rather to the uncontrollable intuition of the investigator and good luck than to the method itself.

While the Stanford-Binet revision did not involve purposive sampling, Neyman's remarks are applicable to the sampling plan. Furthermore, there is internal evidence in the results of the Stanford-Binet revision which indicates that, in spite of the purposive attempt to secure a "representative" sample, the Stanford-Binet revision sample actually produced a larger sampling error than would have resulted from random sampling of clusters.

Table 3 below gives the number of cases from each of the communities included in the Stanford-Binet sample. It should be noted that 37 percent of the "urban" cases were drawn from San Francisco and 56 percent of the "suburban" cases came from two California communities. This means that 975 cases or 34 percent of the total sample were from California. In addition a disproportionately large number of the "rural" cases (41 percent) came from one community in Vermont. It would, of course, be possible to obtain clusterings in two states as marked as those shown by a random sampling of communities, but the probability of such an outcome is extremely small. It is almost certain that a random (or stratified random) sampling of communities would have given a better geographic distribution (and undoubtedly a lower sampling error) than was actually obtained. This fact is also pointed out by McNemar (3) who expresses "skepticism concerning the representativeness of these communities."

It should also be remembered that the authors of the Stanford-Binet revision felt very definitely that their results contained a substantial bias. As noted above, the primary objective of the revision was to obtain a scale giving average I.Q.'s of 100 for each chronological age group. Terman and Merrill (6, p. 23) note that the mean I.Q.'s for their age groups run "slightly above 100" and state that this "is the result of intentional adjustment to allow for the somewhat inadequate sampling of subjects in the lower occupational classes." McNemar (3, p. 20) states:

> The fact that the means in Tables 1 and 2 are above 100 should not lead the reader to the erroneous conclusion that the average I.Q. for the population now exceeds 100. The excess here observed is in the proper direction to allow for known bias in our age samplings. When an adjustment is made for bias in occupational status, the age means approach nearer 100, and a further adjustment for inadequate rural representation would tend to bring the values still closer to 100.

Table 6 on p. 36 of Terman and Merrill (6) gives average I.Q.'s for each age group "adjusted for 1930 Census frequencies of Occupational groupings." These averages still show substantial bias, all means except those for ages 4 and 5½ being over 100 and seven age groups having average I.Q.'s over 103. The effect of rural-urban biasing influences is not presented.

Since the method of correcting for bias is not stated, the effect of these corrections on the mean square errors of the sample results cannot be determined. It is probably not possible to make this determination in any event since the presence or absence of biases in occupational or

rural-urban distributions cannot by themselves tell us whether an I.Q. distribution is biased or unbiased and correcting for rural-urban or occupational biases may have very little effect (or even an unfavorable effect) upon I.Q. biases.

In any event, the original sample means of the Stanford-Binet revision contain substantial biases if the true population means are 100. These are shown by the figures in Table 1.*

TABLE 1

AVERAGE I.Q.'S BY AGE GROUPS FOR THE STANFORD-BINET REVISION SAMPLE

	Age Groups			
	2½–5½	*6–13*	*14–18*	*All Cases*
From L—Mean	106.58	103.22	103.03	104.00
Form M—Mean	106.42	103.96	103.32	104.43
Number of Cases	728	1623	619	2970

In view of the probable biases and the considerations with regard to the sampling method presented above, it is not at all unreasonable to assume that the Stanford-Binet revision sampling involved a larger standard error of the mean than would random selection of communities with equal probability. Even if this is not the case, the subsampling involved should account for an increase in sampling error over a design in which there was no subsampling.

On the basis of the above discussion, the standard error of random selection of communities with equal probability and no subsampling gives us minimum values for the standard errors of the Stanford-Binet sample means. To estimate these errors, we shall assume that the number of cases actually sampled in each community was the total eligible population in that community. (As noted above, assuming that the community population was larger than the number sampled would lead to a larger estimate of the standard error.)

* The data are from McNemar (3) Tables 1 and 2. There are minor differences between the results presented by McNemar and those presented by Terman and Merrill (apparently due to inclusion of some subjects in some of the distributions and their omission in other distributions). The differences are minor and do not affect the present analysis.

The data in this and in the two subsequent tables are reproduced with the permission of Houghton Mifflin Co., the publishers of McNemar's *The Revision of the Stanford-Binet Scale.*

The Standard Error for Cluster Sampling

The standard error for the type of sampling described (i.e. "cluster sampling") is given by:

$$\sigma_{\bar{x}'}^2 = \frac{M - m}{(M - 1)m} \frac{\sum_i^M N_i^2(\bar{x}_i - \bar{x})^2}{M\overline{N}^2} . \qquad [1]$$

Or, when we estimate $\sigma_{\bar{x}'}$ from the sample, the estimated standard error is given by:

$$s_{\bar{x}'}^2 = \frac{M - m}{Mm} \frac{\sum_i^m N_i^2(\bar{x}_i - \bar{x}')^2}{(m - 1)(\overline{N}')^2} \qquad [2]$$

where M = the total number of clusters (communities) in the population
m = the number of communities sampled
N_i = the number of individuals (eligible for the population) in the i-th cluster
$\bar{x}_i$ = the mean I.Q. for the N_i individuals in the i-th cluster

$\bar{x} = \dfrac{\sum_i^M N_i\bar{x}_i}{\sum_i^M N_i}$ = the mean I.Q. of the population. (The aim of the sample is to estimate $\bar{x}$.)

$\overline{N} = \dfrac{\sum_i^M N_i}{M}$ = the average number of individuals per cluster in the population.

$\bar{x}' = \dfrac{\sum_i^m N_i\bar{x}_i}{\sum_i^m N_i}$ = the mean I.Q. of the sample. (We are using this as our estimate of $\bar{x}$ and $s_{\bar{x}'}$ is the estimated standard error of this sample mean.)

$\overline{N}' = \dfrac{\sum_i^m N_i}{m}$ = the average number of individuals per cluster in the sample.

To determine $s_{\bar{x}'}$ exactly we need to know M, the number of communities (clusters) in the population. While M is not known with any precision, we can be quite certain that it is large and that it is much

larger than m (at least 100 times as great). Consequently, we can, without appreciable error, take $M-m/M$ equal to 1. With this substitution the square of the standard error is approximately equal to:

$$s_{\bar{x}'}^2 = \frac{\sum_i^m N_i^2(\bar{x}_i - \bar{x}')^2}{m(m-1)(\overline{N}')^2} = \frac{m}{m-1} \frac{\sum_i^m N_i^2(\bar{x}_i - \bar{x}')^2}{\left(\sum_i^m N_i\right)^2}. \qquad [3]$$

All the data required for Equation [3] can be obtained from the sample. Unfortunately, not all of the sample data are available in published form. Since we shall have to rely on published data, some further approximations (described below) are necessary. The approximations also act to reduce our estimate of the standard error.

McNemar (3) gives, as Table 9, information on the average I.Q.'s for children in "urban," "suburban" and "rural" communities by age groups. This table, plus data for the entire group in the age range 2 to 18, is presented in Table 2. The data for the entire group were calculated from the information given for the three age groups.

TABLE 2

I.Q. DATA FOR URBAN, SUBURBAN AND RURAL CHILDREN*

	Urban	*Suburban*	*Rural*	*Total*
2–5½ Year-Olds				
Number	354	158	144	656
Mean	106.3	105.0	100.6	104.7
S.D.	15.7	16.1	15.4	15.9
6-14 Year-Olds				
Number	864	537	422	1823
Mean	105.8	104.5	95.4	103.0
S.D.	14.7	16.8	15.5	16.1
15–18 Year-Olds				
Number	204	112	103	419
Mean	107.9	106.9	95.7	104.6
S.D.	16.5	15.7	15.9	16.9
All Ages (2–18)				
Number	1422	807	669	2898
Mean	106.2	104.9	96.6	103.6
S.D.	15.2	16.5	15.7	16.2

* Denver 2- to 5½-year-olds are excluded.

To determine $s_{\bar{x}'}$ we shall take for our values of $\bar{x}_i$ (the mean I.Q. in each community): (a) the average I.Q. for urban children for each of

the communities classified as "urban by McNemar; (b) the average I.Q. for suburban children for each of the communities classified as "suburban" and (c) the average I.Q. for rural children for each of the communities classified as "rural." This approximation ignores all variations between communities within the urban, suburban and rural groups of communities. As a result the values of $s_{\bar{x}'}$ which we obtain should be equal to or less than the values which would be obtained if we knew the means of each of the sampled communities.*

There are some uncertainties in the published data concerning the values of m and N_i. As noted above, Terman and Merrill (6) state that 17 communities were sampled in 11 states. This would give $m=17$. However on pp. 36–37, McNemar (3) lists the communities sampled and the number of subjects in each community. McNemar lists 7 urban communities. He also lists 3 suburban communities and states that, in the suburban group, there were "four small communities just out of Kansas City in Johnson County Kansas, with 199 cases drawn from Westwood View, Hickory Grove, Roseland, and Shawnee Mission schools." For the rural communities, McNemar states:

> The samplings from rural communities include 85 from Mount Washington School, Bullitt County, and Liberty School, Oldham County, Kentucky. A total of 152 were drawn from the following districts of Indiana: Prather School, Charlestown schools and Morgan Township School in Harrison County and Galena School in Floyd County. A farming region at Bloomington, Minnesota, supplied 92 cases; the farming and small village community of Randolph, Vermont, provided 275; and 65 subjects were secured in the vicinity of Atlee, Virginia. We have already expressed some skepticism concerning the representativeness of these communities.

From this statement, it is difficult to determine the exact number of "rural communities" involved. At a minimum, there appear to be 8 (assuming that schools in different counties represent different communities). If we also consider the four schools in the "suburban" part of Johnson County, Kansas, to be one community, McNemar's listing gives a count of 19 communities vs. Terman and Merrill's 17. The difference appears to be one in the definition of community. In terms of independently selected areas, Terman and Merrill's "17 communities" is probably more nearly correct. However, the data in Table 2 are based on McNemar's classification. It appears desirable to adopt a compromise, counting as communities the cities and towns listed by McNemar

* This statement cannot be made absolutely since, under certain circumstances, it may be incorrect. However, it is a fairly safe statement since the circumstances which would give a higher standard error through substituting group averages for individual averages are extremely unusual.

plus any schools in separate counties. This is the same basis we used in getting the count of 19 communities mentioned above. Since the count of independently sampled communities is probably Terman and Merrill's figure of 17, this handling of the problem operates in the same direction as the other approximations previously made.

The difficulty in determining N_i occurs in the cases where McNemar gives one figure for the number sampled in two different counties (e.g.,

TABLE 3

NUMBER OF CASES SAMPLED IN EACH COMMUNITY AND ESTIMATED AGE DISTRIBUTION

Communities	*2–5½ Year-Olds*	*6–14 Year-Olds*	*15–18 Year-Olds*	*All Ages (2–18)*
Urban				
1. Denver, Col.	28	67	16	111
2. Minneapolis, Minn.	46	111	26	183
3. New York, N. Y.	12	29	7	48
4. Reno, Nev.	28	68	16	112
5. Richmond, Va.	46	114	27	187
6. San Antonio, Texas	63	155	36	254
7. San Francisco, Calif.	131	320	76	527
Suburban				
8. White Plains, N. Y.	31	107	22	160
9. Redwood City, Calif.	26	89	19	134
10. Los Gatos, Calif.	62	209	43	314
11. Johnson County, Kan.	39	132	28	199
Rural				
12. Bullit County, Ky.	9	27	7	43
13. Oldham County, Ky.	9	27	6	42
14. Clark County, Ind.	11	32	8	51
15. Harrison County, Ind.	11	32	8	51
16. Floyd County, Ind.	11	31	8	50
17. Bloomington, Minn.	20	58	14	92
18. Randolph, Vt.	59	174	42	275
19. Atlee, Va.	14	41	10	65

85 cases from Bullitt County, Kentucky and Oldham County, Kentucky). These cases can be handled by distributing the cases equally among the counties involved. This adjustment also operates to reduce the estimated standard error. A further approximation is necessary to get standard errors for the means of each of the three age groups in Table 2. McNemar gives only the total number of cases in each community and does not give the distribution of these cases among the age groups. To estimate the standard errors for the separate age groups, the number

of cases for each of the communities was distributed by age proportionately to the age distribution in the class (urban, suburban or rural) in which the community falls. The number of cases in each community shown by McNemar and the calculated distribution of these cases by age groups is shown in Table 3. This adjustment affects only the estimates of the standard errors of the age group averages and not the standard error for the entire group aged 2 to 18.

Comparison of Cluster Sampling Error with Unrestricted Random Sampling Error

With all the adjustments reducing the standard error which have been made, it may seem surprising that we have any error left. However, a fairly substantial amount of sampling error remains. Table 4 shows the standard errors of the mean I.Q. calculated as described above (using Equation 3) compared with the standard error obtained by the formula usually used in psychological research studies, i.e.:

$$\sigma_{\bar{x}'}^2 = \frac{\sigma^2}{N'} \qquad [4]$$

where

$$\sigma^2 = \frac{\sum_{i}^{m} \sum_{j}^{N_i} (x_{ij} - \bar{x}')^2}{N'} \qquad [5]$$

and

$$N' = \sum_{i}^{m} N_i. \qquad [6]$$

In Equations [4], [5] and [6], x_{ij} stands for the value (I.Q.) of the jth individual in the ith cluster (community) and the other symbols have the meanings previously defined. Equation [4] represents the standard error of the mean of a sample drawn by unrestricted random sampling from an infinite population (i.e. a sample drawn so that the probability of drawing any observation in the population is equal to and independent of the probability of drawing each of the other observations).

It will be seen from Table 4 that the absolute values of the standard errors calculated by Equation [3] are not large. There is a sampling error of only 1 per cent in the average I.Q. for the entire group of 2,898 cases. However, a very substantial difference exists between the stand-

ard error by Equation [3] and the standard error by Equation [4]. If we apply Equation [4] to determine the standard error of the mean of a cluster sample, it is obvious that we shall be very far from the correct value (in this case we would get an error which is less than one-third of the correct figure).

This fact is extremely important in applying tests of significance to differences of sample means. For example, suppose we took a sample

TABLE 4

ESTIMATED STANDARD ERRORS OF THE MEAN I.Q.'S FOR CLUSTER SAMPLING AND UNRESTRICTED RANDOM SAMPLING

Age Groups	*Standard Errors*		*Ratio of S.E. of Cluster Sampling to S.E. of Random Sampling*
	Cluster Sampling	*Unrestricted Random Sampling*	
2–5½ years	.60	.62	.97
6–14 years	1.09	.38	2.89
15–18 years	1.35	.82	1.63
All Ages (2–18)	1.01	.30	3.36

of 900 children aged 2–18 (by a method which was actually random) from some city or other population subgroup. Assume that this sample gives us an average I.Q. of 105.7 on the revised Stanford-Binet and our sample has a standard deviation of 18, so that the standard error of the mean (using, quite properly, Equation [4]) is .60. Our group has a mean 2.1 points above the average of 103.6 for the Stanford-Binet revision sample shown in Table 2. We want to know whether this difference is significant. If we assume unrestricted random sampling of the Stanford-Binet revision sample, we would use .30 (see Table 4) as the standard error of the revision sample mean. This would give us .67 as the standard error of the difference of 2.1 and our difference would be 3.1 times its standard error. We would undoubtedly consider this a significant difference. Actually, the standard error of the mean of the revision sample is at least 1.01, which makes the standard error of the difference 1.17. The difference is actually only 1.8 times its standard error and can hardly be considered significant.

The sample used for the Stanford-Binet revision is not an extreme case of the error which can be made by applying formulae based on unrestricted random sampling to data obtained by cluster sampling. The sampling for the Stanford-Binet revision did involve testing individuals from several communities and the standard error for cluster sampling is

only 3 times the error for random sampling of the same number of individuals. Many studies use data from one or two groups (e.g.) elementary psychology classes in two neighboring colleges) to draw conclusions about the whole population (all college students or even all human beings). In this case the standard error obtained from Equation [3] may be 50 to 100 times greater than that obtained from Equation [4]. Use of the "correct" formula ("correct" if we have used a random process for drawing our groups) will make supposedly significant differences vanish more rapidly than a quart of ice cream at a children's party.

Intraclass Correlation

The reason for the difference between the standard error for unrestricted random sampling and that for cluster sampling is to be found in the fact that individuals are not sampled independently in cluster sampling. If we consider samples of equal size from the same population, the standard error of the mean in unrestricted random sampling is multiplied by approximately $(1+\overline{N}\rho)$ when we use cluster sampling. Here $\overline{N}$ is the average size of our clusters and ρ is the intraclass correlation (a measure of the extent to which individuals within a cluster resemble, or are "correlated" with, each other). The intraclass correlation usually ranges from 0 to +1 (although it can be negative). It can be seen that even very small values of the intraclass correlation (say, .01) can have a very substantial effect on the standard error of a mean in cluster sampling if the clusters are moderately large ($\overline{N}=100$ or more). As a matter of fact, the estimated intraclass correlation for the entire sample (all individuals aged 2 to 18) used in the Stanford-Binet revision is only .08. A recent paper by Walsh (7) gives some of the probability considerations involved in tests of significance when intraclass correlation is present.

There is one feature of Table 4 which may arouse some interest. This is the fact that the estimated standard error (using Equation [3]) of the mean I.Q. is larger for the group aged 6–14 years than for the group aged 2–5½ years. This is, of course, contrary to what we would expect from the description given of the sampling process. To some extent this peculiarity results from our ignoring subsampling in calculating the standard errors. Consideration of subsampling variation would probably increase the standard errors somewhat and would probably increase the standard error more for the group aged 2–5½ years than for the group aged 6–14 years (since there are fewer of the younger children). As a matter of fact, inclusion of subsampling variation might

double the standard error for the mean I.Q. of the group aged 2–5½ but would probably not increase the standard error of the group aged 6–14 more than 10 per cent

Actually Table 4 shows a lower standard error of cluster sampling for the group aged 2–5½ years than for the group aged 6–14 years because there is less variation among the average I.Q.'s of the urban, suburban and rural children for the younger group. This fact may be due to some basic relation between I.Q. variability and age. For example, McNemar (3) gives a table for adjusting I.Q.'s for differing standard deviation of the I.Q. at various ages. He bases this table on the differences actually found in the sample.

Another explanation of the differences in variability between age groups is to be found in the selective nature of the sampling for the Stanford-Binet revision. Selective sampling seems to have been particularly important in the pre-school group. In another article (4), the present writer points out some effects of selective sampling on group means and also notes that selective sampling will usually affect the standard deviation also. It would be very unwise to hypothesize about the difference between age groups shown in Table 4 unless we had much more information about the sampling probabilities.

This article has used the Stanford-Binet only as an illustration of the dangers of ignoring the intraclass correlation when we are dealing with a cluster type of sampling. In view of the qualifications placed on our analysis, it is not possible to draw any conclusions about the reliability or unreliability of the revised Stanford-Binet as a measuring instrument. There may be good reasons for supposing that the precision of the revised Stanford-Binet is considerably less than many of its users assume. From the sampling standpoint, the sample design used in the revision of the Stanford-Binet was a non-measurable design and there is no way of telling how "bad" or "good" the results were. It has been suggested that the sampling errors shown in Table 4 are probably minimum figures. However, the results do offer a possibility of improving the sampling design in the event that the Stanford-Binet is revised again in the future. An error of 1 I.Q. point in the average I.Q. may not be too serious. If this is the case, the biases in the Stanford-Binet average I.Q.'s could probably be removed by using a sound sample design without any need for an increase in either the number of communities covered or the number of subjects tested. If greater accuracy than a mean correct within 1 per cent is considered necessary or desirable, this could probably be achieved by increasing the number of communities sampled without increasing to any great extent the total

number of subjects tested. As a matter of fact, increasing the number of subjects tested would probably add very little to the accuracy of the final results (at least for the age group 6–14 years). The standard error of a mean in cluster sampling decreases (approximately) in proportion to the square root of the number of clusters sampled. The standard error shown in Table 4 for unrestricted random sampling is .3 of an I.Q. point. The standard error for cluster sampling is 3.36 times this value. Therefore, to get a standard error of .3 using cluster sampling, we would need about 11 times as many communities or about 200 communities. This estimate of the number of communities required is, of necessity, unreliable, since we were forced to estimate our standard errors from a sampling plan which is actually non-measurable, and measuring the non-measurable puts an obvious strain on epistemology.

In designing a sampling plan for a revision of the Stanford-Binet recent developments in sampling theory and practice can be used to increase accuracy without increase in survey costs. The reader's attention is directed particularly to the work of Hansen and Hurwitz (1) in this field. Using the techniques developed by Hansen and Hurwitz, persons revising the Stanford-Binet would probably get satisfactory precision from a well-designed sample of 25 to 100 communities with only a very small increase (if any) in the total number of cases tested.

Summary

This article stresses the dangers of ignoring the intraclass correlation of the population when "cluster" methods of sampling are used. The increase in sampling error resulting from cluster sampling is demonstrated by an analysis of the results of the sample used in the revision of the Stanford-Binet. This sample actually yields "non-measurable" results, i.e. results which do not permit determination of the standard error of the sample mean. However, it is estimated that the standard error of the average I.Q. of this sample is at least 3 times the error which would be calculated by the use of the formula for unrestricted random sampling from an infinite population. The latter formula is the one familiar to psychologists and the one usually used by them regardless of the type of sampling involved. The illustration indicates that very substantial errors may result from this practice and that many results will be considered statistically significant where such a conclusion is entirely unwarranted.

BIBLIOGRAPHY

1. HANSEN, M. H., & HURWITZ, W. N. On the theory of sampling from finite populations. *Ann. math. Statist.*, 1943, **14**, 333–362.
2. MCNEMAR, QUINN. Sampling in psychological research. *Psychol. Bull.*, 1940, **37**, 331–365.
3. MCNEMAR, QUINN. *The revision of the Stanford-Binet Scale.* Boston: Houghton Mifflin, 1942.
4. MARKS, ELI S. Selective sampling in psychological research. *Psychol. Bull.*, 1947, **44**, 267–275.
5. NEYMAN, JERZY. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *J. roy. statist. Soc.*, n. s., 1934, **97**, 558–606.
6. TERMAN, L. M., & MERRILL, MAUD A. *Measuring intelligence.* Boston: Houghton Mifflin, 1937.
7. WALSH, J. E. Concerning the effect of intraclass correlation on certain significance tests. *Ann. math. Statist.*, 1947, **18**, 88–96.

APPENDIX

Although the formula for the standard error of the mean for cluster sampling is not new, psychologists are generally unfamiliar with it. The derivation of this formula is, therefore, presented below. The z transformation will be found useful in deriving standard errors for more complicated designs (e.g. designs using stratification, subsampling, differential sampling probabilities, etc.).

Equation [2] gives the mean square error (square of the standard error) of the mean of a cluster sample as:

$$s_{\bar{x}'}^2 = \frac{M - m}{Mm} \frac{\sum_i^m N_i^2(\bar{x}_i - \bar{x}')^2}{(m-1)(\overline{N}')^2}.$$

The mean of the sample is:

$$\bar{x}' = \frac{\sum_i^m N_i \bar{x}_i}{\sum_i^m N_i}.$$

M, m, N_i, $\bar{x}_i$, $\bar{x}'$, $\overline{N}'$, $\bar{x}$ and $\overline{N}$ are defined on p. 420 It is also convenient to define:

x_{ij} = value for jth individual in ith cluster

$x_i = \sum_j^{N_i} x_{ij}$ = sum of the values for all individuals in ith cluster.

From their definitions, it can be seen that:

$$\bar{x}_i = \frac{x_i}{N_i}$$

and, therefore

$$\bar{x}' = \frac{\sum_i x_i}{\sum_i^m N_i}.$$

$\bar{x}'$ can be treated as a ratio of two linear functions of the sample observations, namely:

$$f(x) = \frac{M}{m}\sum_i^m x_i = \frac{M}{m}\sum_i^m \sum_j^{N_i} x_{ij}$$

$$f(N) = \frac{M}{m}\sum_i^m N_i = \frac{M}{m}\sum_i^m \sum_j^{N_i} 1.$$

In deriving $s_{\bar{x}'}^2$, it will be useful to prove the following theorem:

Theorem: If we have a sample estimate:

$$r' = \frac{f(x)}{f(y)}$$

where $f(x)$ and $f(y)$ are linear functions of the sample observations x_h and $y_h (h = 1, 2 \cdots n)$ and if:

$$x = Ef(x), \qquad y = Ef(y), \qquad z_h = \frac{x_h}{x} - \frac{y_h}{y}$$

then:

$$\sigma_{r'}^2 = \left(\frac{x}{y} - r\right)^2 + \left(\frac{x}{y}\right)^2 \sigma_{f(z)}^2$$

where $\sigma_{r'}^2$ is the mean square error of r' and r is the population parameter of which r' is an estimate.

Proof: When we have a sample estimate $r' = f(x)/f(y)$, the mean square error of r' can be found by: (a) expanding r' as a Taylor series around x and y (the expected values of $f(x)$ and $f(y)$); (b) subtracting r (the true value of r' for the entire population) from both sides of the equation; (c) squaring both sides of the resulting equation and (d) taking the expected value of the resultant. If we ignore, in our Taylor series, terms involving partial derivatives higher than the first, the result of this operation will be:

$$\sigma_{r'}^2 = E(r' - r)^2$$

$$= \left(\frac{x}{y} - r\right)^2 + \left(\frac{x}{y}\right)^2 \left(\frac{\sigma_{f(x)}^2}{x^2} + \frac{\sigma_{f(y)}^2}{y^2} - 2\frac{\sigma_{f(x)f(y)}}{xy}\right). \qquad [7]$$

If we let

$$z_h = \frac{x_h}{x} - \frac{y_h}{y}$$

and if f is a linear function, then:

$$f(z) = \frac{f(x)}{x} - \frac{f(y)}{y} \qquad [8]$$

and

$$Ef(z) = \frac{Ef(x)}{x} - \frac{Ef(y)}{y} = \frac{x}{x} - \frac{y}{y} = 0 \qquad [9]$$

$$\sigma_{f(z)}^2 = \frac{\sigma_{f(x)}^2}{x^2} + \frac{\sigma_{f(y)}^2}{y^2} - \frac{2\sigma_{f(x)f(y)}}{xy}. \qquad [10]$$

Therefore:

$$\sigma_{r'}^2 = \left(\frac{\bar{x}}{y} - r\right)^2 + \left(\frac{x}{y}\right)^2 \sigma_{f(z)}^2. \qquad [11]$$

The above theorem can be applied to derive the mean square error of $\bar{x}'$, as follows:

$$\bar{x}' = \frac{f(x)}{f(N)}$$

where

$$f(x) = \frac{M}{m}\sum_i^m x_i \quad \text{or} \quad f(x) = \frac{M}{m}\sum_i^m \sum_j^{N_i} x_{ij}$$

$$f(N) = \frac{M}{m}\sum_i^m N_i \qquad f(N) = \frac{M}{m}\sum_i^m \sum_j^{N_i} N_{ij} \quad \text{where } N_{ij} \equiv 1$$

and

$$Ef(x) = \sum_i^M x_i = x$$

$$Ef(N) = \sum\sum N_i = N.$$

Let

$$z_i = \frac{x_i}{x} - \frac{N_i}{N} \quad \text{or}^* \quad z_{ij} = \frac{x_{ij}}{x} - \frac{1}{N}$$

$$f(z) = \frac{M}{m}\sum_i^m z_i = \frac{\frac{M}{m}\sum_i^m x_i}{x} - \frac{\frac{M}{m}\sum_i^m N_i}{N}.$$

By Equation [11]:

$$\sigma_{\bar{x}'}^2 = \left(\frac{x}{N} - \bar{x}\right)^2 + \left(\frac{x}{N}\right)^2 \sigma_{f(z)}^2 \qquad [12]$$

where $\bar{x}$ is the population parameter estimated by $\bar{x}'$ and is:

$$\bar{x} = \frac{\sum_i^M x_i}{\sum_i^M N_i} = \frac{x}{N}.$$

Therefore:

$$\sigma_{\bar{x}'}^2 = \left(\frac{x}{N}\right)^2 \sigma_{f(z)}^2. \qquad [13]$$

Since $f(z)$ is M/m times a sum of the sample values z_i:

$$\sigma_{f(z)}^2 = \frac{M^2(M-m)}{(M-1)m} \frac{\sum_i^M (z_i - \bar{z})^2}{M} \qquad [14]$$

where

$$\bar{z} = \frac{\sum_i^M z_i}{M} = Ef(z) = 0. \qquad [15]$$

An unbiased estimate of $\sigma_{f(z)}^2$ from the sample is:

$$s_{f(z)}^2 = \frac{M^2(M-m)}{Mm} \frac{\sum_i^m (z_i - \bar{z}')^2}{m-1} \qquad (16)$$

* The result is the same whether the z transformation is applied to the cluster totals or the individual observations.

where

$$\bar{z}' = \frac{\sum_i^m z_i}{m} = \frac{\sum_i^m x_i}{mX} - \frac{\sum_i^m N_i}{mN}.$$

From Equations [13], [14], and [16] we have:

$$\sigma_{\bar{x}'}^2 = \left(\frac{x}{N}\right)^2 \frac{M^2(M-m)}{(M-1)m} \frac{\sum_i^M (z_i - \bar{z})^2}{M} \qquad [17]$$

and

$$s_{\bar{x}'}^2 = \left(\frac{x}{N}\right)^2 \frac{M^2(M-m)}{Mm} \frac{\sum_i^m (z_i - \bar{z}')^2}{m-1}. \qquad [18]$$

In Equation [17] we substitute the values:

$$z_i = \frac{x_i}{x} - \frac{N_i}{N}, \qquad \bar{z} = 0, \qquad \overline{N} = \frac{\sum_i^M N_i}{M}$$

and get:

$$\sigma_{\bar{x}'}^2 = \frac{(M-m)}{(M-1)m} \frac{\sum_i^M \left(x_i - N_i \frac{x}{N}\right)^2}{M\overline{N}^2} \qquad [19]$$

or

$$\sigma_{\bar{x}'}^2 = \frac{M-m}{(M_s-1)m} \frac{\sum_i^M N_i^2(\bar{x}_i - \bar{x})^2}{M\overline{N}^2}. \qquad [20]$$

We make the same substitutions in Equation [18] and also substitute for $\bar{x}$ and $\overline{N}$ the sample estimates:

$$\bar{x}' = \frac{f(x)}{f(N)} = \frac{\sum_i^m x_i}{\sum^m N_i}$$

and

$$\overline{N}' = \frac{f(N)}{M} = \frac{\sum_i^m N_i}{m}.$$

This gives:

$$s_{\bar{x}'}^2 = \frac{M - m}{Mm} \frac{\sum_i^m N_i^2(\bar{x}_i - \bar{x}')^2}{(m-1)(\overline{N}')^2} \qquad [21]$$

or

$$s_{\bar{x}'}^2 = \frac{M - m}{M} \frac{m}{m-1} \frac{\sum_i^m N_i^2(\bar{x}_i - \bar{x}')^2}{\left(\sum_i^m N_i\right)^2}. \qquad [22]$$

In some cases, cluster sampling may introduce a substantial bias into the sample standard deviation (when the sample S.D. is used as an estimate of the population S.D.). This bias will be practically eliminated by use of the estimate:

$$s_s^2 = \sigma_s^2 + s_{\bar{x}'}^2 \qquad [23]$$

where σ_s is the sample S.D. and s_s is an estimate of the population S.D.

Equation [23] can also be used for estimating the population S.D. from a sample with unrestricted random sampling.

ILLUMINATION STANDARDS FOR EFFECTIVE AND EASY SEEING

MILES A. TINKER
University of Minnesota

The problem of artificial illumination is of primary importance in all inside working environments. To maintain healthful and efficient functioning of the eyes, it is necessary to provide adequate lighting. Unquestionably, proper illumination contributes much to comfort and efficiency in activities of daily life. Working under faulty illumination frequently results in eyestrain which tends to be accompanied by reflex functional disturbances of other organs.

During recent years a "lighting consciousness" has been forced upon a large portion of the population, particularly upon those who do considerable visual work under artificial light and upon those who must decide upon the illumination requirements of schools, offices, factories and other situations where visual work is to be performed. Although interest in lighting has been stimulated by popular articles, advertisements, and "educational pamphlets"—as well as by reports written by educators and medical men—the more fundamental information has appeared as experimental reports in scientific publications. The result of exposure to this material is a keen interest in illumination and a sincere desire on the part of the public for sound information concerning hygienic lighting. The natural tendency is to consult pamphlets on recommended practice when lighting specifications are needed for a particular situation. Frequently, the applied psychologist will be called upon to furnish advice on proper illumination. In many instances he will be asked to evaluate the materials presented in the recommended practices. Consequently, the applied psychologist should be informed concerning the adequacy of the data from which the lighting specifications in the recommendations are derived.

The first code on lighting was issued by the Illuminating Engineering Society in 1915. In the more recent publications, the codes are known as Recommended Practice of Home Lighting, of Office Lighting, etc. These pamphlets have been prepared by the Illuminating Engineering Society either alone or jointly with the American Institute of Architects, usually under the rules of procedure of the American Standards Association. Although the American Psychological Association has been in existence for over 50 years, and even though applied psychologists have been interested in the field and have been making experimental contributions to the hygiene of vision for over 40 years, neither psy-

chology nor psychologists are represented in the group specifying recommended practices. Furthermore, a large body of psychological literature has been ignored, either because the illuminating engineers were not familiar with it or because they chose not to use it. The result has been an emphasis upon the engineering aspects of lighting with inadequate attention to certain psychological factors. More recently there has been some attempt to consider more of the psychological factors. Perhaps because engineers lack a psychological background, interpretations are frequently erroneous. Probably the most satisfactory approach to hygienic lighting could be achieved by coordinating the contributions of engineers, physiologists, and psychologists.

Recent editions of recommended practices reveal an increased emphasis upon control of direct and reflected glare, brightness contrast, and the diffusion or distribution of light. The tendency to specify relatively very intense light for many visual tasks is prominent. The purpose of this paper is to present a critical examination of the specifications in the more recent editions of recommended practices and to scrutinize some of the data from which the recommendations were derived.

Spectral Quality of Light

In general, spectral quality of light receives adequate treatment in recommended practices (**35, 36, 37, 38**). It is stated that with equal foot candles of illumination, variations in color quality of light found in common illuminants have little or no effect upon the visual discrimination involved. When color is to be discriminated, it should be viewed under as close an approximation of daylight as possible. Luckiesh (**10**) has a valuable discussion of light and color.

Quality of Lighting

Recommendations (**35, 36, 37, 38**) concerning control of glare, diffusion, direction and distribution of light, light reflection value, and effects of finishes on ceilings and wall are ordinarily quite satisfactory. Visual discrimination is improved by moving the glare source away from the line of vision and by reducing the brightness of the light source and the amount of light emitted by the light source toward the eye. Brightness of luminaires should be low in value. High brightness contrasts within the field of vision should be avoided whether on the work surface or in other parts of the visual field. Proper diffusion of light helps to eliminate undesirable shadows. Purely local lighting, therefore, is unsatisfactory. Since the reflection factors of objects in the visual environment play an important role in illumination, the finish of ceilings, walls,

floors and furnishings is important. These surfaces should provide reflecting surfaces to help spread the light about the room. Furthermore, they should be such that undesirable brightness contrast does not occur within the field of vision. Shiny or glossy finishes should be avoided to prevent specular glare.

In the recommended practices, informative discussions on classification of lighting systems are usually included. Also illustrations of fixtures and installations are sometimes given. Some attention is given to daylight illumination and the need of coordinating artificial with daylight lighting.

Intensity of Illumination

Intensity of illumination receives by far the greatest emphasis in specifications. With each revision of a lighting code prepared by illuminating engineers, the foot candle recommendations for a given situation rise. One may well question whether this trend has a scientific basis, or whether the consumer has been educated to accept the higher intensities. In 1934, Luckiesh and Moss (11) presented general recommendations which they considered to be very conservative. These are repeated with slight changes in Luckiesh's 1944 book (10). He adds that these are inadequate in many cases where hundreds and even thousands of foot candles of light are desirable. Examination of the recommended practices of lighting reveals that, for the most part, they are based upon researches done and interpretations made by Luckiesh and his co-workers, or upon researches inspired by them. Let us turn first, therefore, to these reports.

In *Light, Vision and Seeing*, Luckiesh (10), and in the *New Science of Seeing*, Luckiesh and Moss (11), make the following foot candle recommendations for common tasks of the work-world:

1. 100 foot candles or more are specified for severe and prolonged visual work. Examples include fine needle work, pen work, engraving and assembly, and discrimination of fine details involving low contrast.
2. 50 to 100 foot candles should be used for proof-reading, difficult reading, watch repairing, and average sewing.
3. 20 to 50 foot candles are listed for such visual tasks as clerical work, ordinary reading and average sewing on light goods.
4. 10 to 20 foot candles are proposed for ordinary reading and sewing on light goods when the task is not prolonged.
5. 5 to 10 foot candles are needed for visual work which is more or less interrupted or casual.
6. 1 to 5 foot candles are sufficient for perceiving large objects.

Luckiesh (10) states that these are minimum foot candle recommendations and that he considers them to be very conservative from the

viewpoint of ease of seeing. Furthermore these foot candles, according to Luckiesh and Moss (11), are far below the intensities of illumination which new knowledge indicates to be ideal.

These recommendations are derived from various sets of data which will be discussed in turn.

Preferences for light intensity. Luckiesh and Moss (11) cite data on preferences for light intensities to support their contentions that high intensities are necessary for adequate seeing. The mean choice was about 100 foot candles but the median was 50 foot candles when up to 1000 foot candles were available. Tinker's analysis (22) of light preference studies indicated that visual adaptation plays an important role in determining the preferences. In an experimental check, Tinker (26) found that when readers were adapted to 8 foot candles, the median choice for comfortable reading was about 12 foot candles. But when adapted to 52 foot candles, the median choice was 52 foot candles. It is obvious that the intensity of illumination to which the reader is adapted plays a dominant role in his illumination preference. The conclusion is, therefore, that preference for illumination intensity is not a satisfactory method for determining the intensity of light needed for efficient visual work.

Visual acuity. Luckiesh and Moss (11) and Luckiesh (10) list visual acuity as a basic factor in reading (and presumably in other visual work). It is true enough that visual discrimination does depend somewhat upon visual acuity. But is visual acuity an adequate criterion for prescribing appropriate lighting? Luckiesh and Moss (13) admit that in many tasks the criterion of visual acuity is relatively inappropriate, e.g. in tasks involving low contrasts. But they point out that for black test objects on a white background, visual acuity improves up to 100 foot candles. As a matter of fact, Lythgoe (15) has shown that under certain conditions of measurement, visual acuity improves up to and beyond 1000 foot candles. Inspection of the data reveal that the knee of the curve of improvement is at about 10 foot candles and that beyond about 20 foot candles the gains are slight. It must be kept in mind that in measuring visual acuity, one is dealing with threshold values. It is highly questionable whether the almost microscopic gains in visual acuity obtained under the high foot candles justify their application to visual tasks where supra-threshold visibility is involved as in most everyday situations. Furthermore, data reveal that the visual acuity curve is practically horizontal from 50 foot candles to the higher levels.

Luckiesh and Moss (11) and Luckiesh (10) cite data on visual acuity for 1, 10, and 100 foot candles only. If they really desired to find the foot candle level beyond which no *practical* gains in visual acuity occur, they should have investigated the range between 10 and 100 foot can-

dles. As shown in Tinker's reviews (**29, 31**), this criticism may be aimed at all the basic data presented by Luckiesh (**10**). In some instances (decrease in heart rate, decrease in convergence reserve of ocular muscles), data for only 1 and 100 foot candles are presented. This procedure is inexcusable in experiments designed to determine how much light intensity is needed for efficient visual work. It appears, then, that visual acuity data are of only slight use for prescribing illumination intensities for visual discrimination in supra-threshold tasks. If accepted, there is no justification for suggesting that more than 40 to 50 foot candles are necessary for adequate discrimination even for tasks that approach threshold discrimination.

Visibility measurements. Luckiesh (**10**) states that "After establishing a standard of visibility or desirable see-level to be attained if possible for all tasks, it is seen that specifications of light and lighting and other aids to seeing can be based upon visibility measurements." The measurements are to be made by the Luckiesh-Moss Visibility Meter. This is a device consisting of two identical circular gradients which are rotated before the eyes to alter the brightness contrast of the object whose visibility is to be measured. It, therefore, reduces the object to threshold visibility. It is the threshold which is measured. Three assumptions are made: (a) Two objects are equal in visibility when both are barely visible, (b) "Two objects are equally above threshold visibility when their visibility has been increased by the same increase" in size, brightness, brightness contrast or time, (c) "The visibility of an object, or degree of supra-threshold visibility, is proportional to the decrease in any one of the fundamental factors necessary to reduce the object to threshold visibility." These assumptions are considered to be *axiomatic* and arguments against them are considered to be *futile*. Nevertheless, since recommended standards are based upon visibility measurements to a large degree, it seems desirable to examine the matter further. Things are not axiomatic just because some one says they are.

Since visibility measurements are in terms of threshold values, they are analogous to visual acuity measurements. They are subject, therefore, to the same criticisms as visual acuity measurements as criteria for prescribing illumination standards.

Luckiesh (**10**) emphasizes foot candles for equal visibility in prescribing illumination intensities. For example, to make newspaper text matter equivalent in visibility to 8 point book type on white paper under 10 foot candles of light, it is necessary to use 30 foot candles. And to make the 1/64″ divisions on a steel scale equal to this visibility level, 180 foot candles are needed. Are these levels of illumination intensity required for efficient and comfortable seeing? Luckiesh (**10**) assumes that this is a conservative standard. On his empirical scale, the 8 point type with 10 foot candles has 48 per cent maximum visibility. (Maxi-

mum visibility is obtained from a test-object whose critical detail has a visual size of 20 minutes; a critical detail of 1 minute is the smallest visible for persons with normal vision under 10 foot candles of light.) But no adequate experimental check is made for performance of these tasks under various levels of illumination. Tinker (27) found that the critical illumination level (the intensity beyond which no further change in reading performance occurs as the intensity is increased) for reading 7 point newspaper type to be approximately 7 foot candles. It is difficult to conceive the need of going above 20 foot candles to provide a margin of safety above the critical level. It is highly probable that an experimental check will reveal that other visual tasks, like discriminating the divisions on a steel scale, do not require the 180 foot candles indicated for efficient vision by the *computations* of Luckiesh. Related to this is the question of comfortable vision. Harrison (8), in discussing the difficulty of using high intensities because of the introduction of glare factors states "Visibility and comfort are two separate factors which do not always overlap completely."

No one will deny that visibility is an important factor in ease of seeing. But to prescribe standards in terms of scores derived from measurements made with the Visibility Meter is open to serious question. The basic data are threshold scores. While the derived scores may appear logical, supra-threshold seeing is not the same phenomenon as threshold seeing. Apparently, as illumination intensity is increased, one soon reaches a level of diminishing returns where further increase is of no practical importance or may introduce harmful factors from the viewpoint of easy and comfortable seeing.

Nervous muscular tension. Luckiesh and Moss (11, 12), place great stress upon the apparent decrease in nervous muscular tension during reading as the illumination intensity is increased from 1 to 10 to 100 foot candles. Tinker's (22) analysis of their data reveals that the method employed to present their results magnifies minute differences so that they appear large. Interpolation shows only gradual changes from 10 to 20 to 25 foot candles and very slight changes from there on to 100 foot candles. The conclusions that high foot candles are needed for ordinary reading is not valid. In a comparable situation, Tinker (25) found that for reading 10 point type, the critical intensity was about 3 foot candles. Below this level, rate of reading was retarded and fatigue increased, but for higher intensities there was no change. For people with normal vision, 10 to 15 foot candles should provide a satisfactory margin of safety for reading legible print.

Frequency of blinking. Another favorite criterion employed by Luckiesh and Moss (11, 12) and Luckiesh (10) as a basis for prescribing illumination intensities for visual work is frequency of blinking. The typical experiment is to measure the rate of involuntary blinking for the first and for the last five minutes for an hour's reading under 1, under

10 and under 100 foot candles of light. They note that the blink rate is greater under the 1 than under 10, and greater under 10 than under 100 foot candles. Therefore, it is concluded that relatively high intensities are desirable for reading. Even if these data are accepted as valid, we do not know where between 10 and 100 foot candles the curve of increased efficiency flattens out since intermediate intensity values were not studied. But there are several sources of information which suggest that blink rate is not a valid criterion of ease of seeing:

1. McFarland, Holway, and Hurvich (18), after a searching analysis of their own extensive experiments and of other studies, state: "A high blink-rate need mean neither an increase in fatigue nor an increase in difficulty of seeing." They conclude that "the rate of blinking can hardly be considered as a valid index of visual fatigue."

2. Tinker (32), in a study that has some bearing on the subject, found that frequency of blinking is an inadequate criterion of readability of print.

3. Bitterman (1), working with 3 and 91 foot candles of light, found that when subjects read for 40 minutes there was no significant difference in rate of blinking. In fact the frequency of blinking was slightly greater under the 91 foot candles. Incidentally, Bitterman also found no significant difference in blink rate for reading large type vs. small type. His results, therefore, indicate that rate of blinking cannot be employed as an index of ease of visual work. Further studies by Bitterman and Soloway (2, 3) showed that frequency of blinking is unrelated to duration of visual work or to the presence of a relatively intense glare source in the visual field. The reports of McNally (19) and MacPherson (16) also cast doubt upon the validity of blinking as an index of ease of seeing.

4. The statistical treatment employed by Luckiesh and Moss (11, 12, 14) upon their data is open to severe criticism. Tinker (28, 29) has questioned the appropriateness of the geometric mean which they employ in most comparisons. The same criticism is raised by Hoffman (9). In a searching analysis, Hoffman also severely criticizes the use of the percentage technique employed by Luckiesh and Moss for presenting data, and for basing conclusions on percentage differences rather than on raw score differences. Percentage scores are notoriously unreliable. Furthermore, if the raw scores are below 100 (as most of them are), percentages magnify the differences. When percentages are used, therefore, the observed differences may be largely an effect of the derivation. Insignificant raw score differences may seem large when put into percentages. For instance, a typical average of 30 blinks during 5 minutes of reading is increased 10 per cent by a change of 3 blinks. Hoffman further points out that work decrement may be a more important variable than illumination changes in the results of Luckiesh and Moss. In general, he found little support for the contention that relatively high intensities are needed for effective and easy seeing.

5. Eames (5) criticizes Luckiesh and Moss (14) for using relatively few subjects in their experiments (including blink rate studies) and for employing "test wise" subjects. As pointed out by Eames, "People who take tests repeatedly in a given field gradually learn what is expected of them" and are un-

intentionally influenced by this knowledge. Results obtained under such conditions cannot be representative of the reactions of the general population.

The accumulated evidence indicates that rate of blinking cannot be accepted as a criterion for specifying intensities of light for visual work.

Decrease in heart rate. Luckiesh (**10**) and Luckiesh and Moss (**11, 12, 14**) cite data on change of heart rate while reading for one hour under 1 foot candle and under 100 foot candles of light. No data are presented for intermediate levels of illumination. It is stated that heart rate decreased 10 per cent under the 1 foot candle and 2 per cent under 100 foot candles. The conclusion was that from the viewpoint of ease of seeing the 100 foot candle level is desirable. An experiment by McFarland, Knehr and Berens (**17**) was designed to check the findings obtained in Luckiesh's laboratory. The results led to the conclusion that "It is questionable whether reliable criteria for determining adequate levels of illumination for tasks such as reading during short periods of time (approximately 2 hours) can be obtained in terms of . . . heart rate . . . ." Another check experiment was carried out by Bitterman (**1**), who recorded heart rate while reading under 3 and under 91 foot candles of light. "The results do not support the conclusions of Luckiesh and Moss with respect to the value of heart rate" as an index "of the ease of visual work." In view of the above evidence we must reject heart rate as a criterion for prescribing illumination intensities for visual work.

Decrease in convergence reserve. Luckiesh and Moss (**11, 12, 14**) and Luckiesh (**10**) cite data on decrease in convergence reserve of ocular muscles after reading for one hour under 1 and under 100 foot candles of light. The decrease was less under the 100 foot candles. No data are given for the range between 10 and 100 foot candles. We do not know, therefore, whether the 100 foot candles is significantly better than such levels as 20 or 30 foot candles.

Visual adaptation. Throughout their writings, Luckiesh and Moss (**10, 11, 12, 14**) emphasize that the eyes evolved under daylight levels of illumination and suggest the desirability of competing with daylight by artificial means. They consistently ignore the fact that the eyes readily adapt to easy and effective seeing over a wide range of illumination intensities.

Summary on intensity of illumination. Examination of the data employed by Luckiesh and Moss as a basis for specifying foot candle levels for visual work reveals a general lack of validity of these results as criteria for ease of seeing. The data from visual acuity, muscular tension and visibility measurements are misinterpreted or misapplied. The blink technique and rate of heart beat must be rejected because of lack of confirmation by independent workers. Furthermore the methods of

statistical analysis employed are frequently at fault. Any science of seeing based upon such an unstable foundation must, therefore, lack validity. Since these data have been the justification for specifying what appear to be excessively high levels of illumination intensity, we must reject such specifications unless justified by valid evidence from new experimentation.

Lighting Codes

School lighting. The American Recommended Practice of School Lighting (**35**) specifies the following minimum foot candles in service: 15 for classrooms, shops and offices; 25 for sewing and drafting rooms; and 30 for sight-saving classes. There is general agreement on the importance of hygienic illumination in reading and study situations. The recommended foot candle levels seem satisfactory in view of research findings other than those cited in the code. There should be, of course, a sound experimental basis for recommendations of this kind. Tinker (**23**) has pointed out that the recommended practice for school lighting is based upon conclusions derived from misinterpreted experimental results. Fortunately, the recommended practice is adequate in spite of inferences from inadequate data.

In a later publication by Sturrock (**21**), the foot candle levels are not in an approved code but are listed as the levels found desirable in the experience of successful business institutions, i.e., good present-day practice. For schools the foot candles listed include: 30 for study halls, class rooms, general laboratories, general manual training; 50 for drawing room, close work in laboratory, sight saving classes; 100 (considered especially low) for close work in manual training, and in sewing rooms. It is obvious to the impartial person who knows the field that these suggestions represent more intense illumination than is necessary for adequate seeing in the school situation. Data summarized by Tinker (**24**) and additional experimental evidence (**25, 27**) indicate that about 15 foot candles are adequate for ordinary schoolroom tasks and that 25 to 30 foot candles are satisfactory for the more severe tasks. Justification for the higher intensities is sought in the discussions of Luckiesh and Moss (**12, 14**) and Luckiesh (**10**). These have been evaluated above.

Office lighting. The Recommended Practice of Office Lighting (**36**) includes the following foot candle levels: 50 for difficult seeing tasks such as accounting, bookkeeping, and drafting; 25 for ordinary seeing tasks such as general office work, private office work, mail rooms; 10 for casual seeing tasks such as reception rooms and washrooms; 5 for simple seeing tasks such as halls and stairways. Considering the se-

verity of the tasks performed by some workers in general offices and special (as accounting) offices, the above recommendations are satisfactory. The 50 foot candles, however, should be considered liberal even for the difficult seeing tasks. The statement that "higher values will contribute greatly to accuracy, speed and ease" cannot be accepted as valid.

Sturrock's (**21**) summary of good present day practice does not deviate markedly from the recommended practice except that typing and prolonged reading of shorthand notes are listed at 50 foot candles and intermittent reading and writing at 30 foot candles. Each of these is about twice what is needed in terms of the visual task. The basis for the higher intensities is in terms of the discussions of Luckiesh and Moss (**12, 14**). The inadequacy of these data has been pointed out above.

Industrial lighting. A wide range of illumination intensities is recommended for various tasks in industry (**37**). Among the higher foot candle recommendations are: *over* 100 foot candles for such operations as extra fine assembly, automobile finishing and inspecting, cutting and sewing dark goods, engraving, proofreading, final inspection of tire casings, grading and sorting tobacco products, and certain inspection work in textiles; 50 to 100 foot candles for such operations as automobile assembly line, glass works inspection, fine inspection, bookkeeping, font assembly-sorting in printing industry, tin plate inspection, and stitching dark leather. With regard to all the recommendations, one is cautioned that the foot candles are minimum operating values and that in almost every instance higher values may be used with greater benefit.

It is stated that the recommendations are taken from a series of studies on the illumination needs of specific industries, or, if not available there, from current good practice. Examination of these studies (listed on page 23 of the report) indicates that in the main they are surveys rather than experiments. Furthermore, there is a lack of adequate descriptions of the survey techniques employed. In a few instances a general description of methods was given. Apparently what happened was first to make a survey of practice. This was followed by some sort of job analysis to determine what had to be discriminated. Then by reference to research studies (such as those reported by Luckiesh and Moss in their books) the intensity level of illumination presumably needed for the specific job was deduced. This method has some virtue provided sound data are referred to, which was not done in these cases. In a few instances it is stated that visibility measurements were made. Occasionally installations to achieve the recommendations

were made, the effect observed and additional modifications made. In no case was there experimental determination of the light intensity needed.

There are no valid experimental data which indicate that more than 50 foot candles are needed even for those practical visual tasks which approach threshold discrimination. Furthermore, as pointed out by Harrison (**8**), visual comfort may decrease under high intensities.

Home lighting. The most recent recommended practice for home lighting (**38**) specifies intensities ranging from 10 foot candles on card tables to 100 and more for sewing on dark goods. Forty foot candles are recommended for such situations as children's study table, kitchen work counter, laundry, and for prolonged reading. There is no valid reason for going above 25 to 30 foot candles for the more severe visual tasks in the home (**24**). Approximately 15 foot candles is adequate for many of these visual tasks. Figure 1 in *Recommended Practice of Home Lighting* (**38**) is misleading. "This chart shows the extent to which occupations and poor seeing conditions leave their mark on eyesight." The implication is that poor illumination causes ocular disability. There are no valid data which indicate this to be so. This chart represents an unjustified form of propaganda.

Present-day practice. Sturrock (**21**) has assembled foot candle levels of illumination which are labeled "good present-day practice." The tables are preceded by a classification (after Luckiesh and Moss) of foot candle needs for visual discrimination of tasks varying in difficulty. The material is apparently designed as a guide but is not necessarily in the form of recommendations. This sort of thing is valuable in many ways. But since it is based to a considerable degree upon the material presented by Luckiesh and Moss (**12, 14**) and by Luckiesh (**10**), the illumination intensities are excessively high in some instances—as 100 foot candles for sewing and proofreading, and 50 foot candles for reading small type and for kitchen counters. It should be pointed out, however, that much of the material is fairly satisfactory.

In general, recommended practice prior to 1940 (**35**) is fairly adequate, but as new codes are issued at later dates the apparent tendency has been to recommend as intense lighting as the traffic will bear. This is justified by referring to the work of engineers (largely Luckiesh and Moss) who state that these high intensities are nevertheless inadequate for easy seeing. As pointed out above, both the experiments and the conclusions which are cited as fundamental are frequently invalid. Furthermore the data are out of line with other independent experimental results.

Visual Factors

Eye disabilities. It is generally accepted that eyes with disabilities, even when corrected by glasses, need brighter light than normal eyes for adequate visual discrimination. Ferree and Rand (6) and Ferree, Rand and Lewis (7) are usually cited as supporting evidence. In the first study (6), it was found that apparent diopters of accommodation increased more for 14 presbyopes than for normal eyes in going from 1 to 5 to 25 foot candles of light. Interpolation indicates that for the normal eyes the curve of improvement shows little rise after about 8–10 foot candles; for the presbyopes, after about 15 foot candles. In addition, *one* myope and *one* presbyope were compared with a normal subject by measuring apparent diopters of accommodation at 13 intensities from 0.5 to 100 foot candles. The curve of efficiency for the normal person improved rapidly to 5 foot candles, then more slowly to about 20 and very gradually thereafter; for the myope there was considerable improvement to about 20 foot candles and little thereafter; for the presbyope there was considerable improvement to about 38 foot candles, and then slower improvement to 100 foot candles. It is of course impossible to generalize from *one case*, but apparently those with eye disabilities need somewhat brighter light than normals for clear seeing. This does not mean that they need 100 foot candles or more, as some people wish to imply.

In the other study (7) Ferree, Rand and Lewis were concerned with distant (20 feet) vision. The visual acuity for 4 presbyopes was compared with acuity for 3 normal people. The presbyopes continued to gain in visual acuity from 25 to 100 foot candles while the normal eye made little gain within this range. Since there is little or no relation between acuity of distant vision and acuity at near vision, these results have no bearing upon visual discrimination at the work surface (desk, work bench, etc.). Furthermore, one should not prescribe illumination for suprathreshold tasks in terms of threshold measurements (visual acuity). There is no evidence from these studies which implies that excessively high foot candles are necessary for those with ordinary visual disabilities. Rather, they suggest a moderate increase for those with corrected vision as compared with normal eyes.

Visual adaptation. It is well established that the eyes readily adapt to easy and effective seeing over a wide range of illumination intensities. This adaptation is rather slow in going from bright to dimmer illumination (for practical purposes, 15–20 minutes) and rapid in going from dim to bright illumination (1-3 minutes). Tinker (25) has demonstrated

that when adaptation is incomplete on shifting to a lower level of illumination, speed of perception is retarded. When adaptation is adequate, however, visual perception in reading is fully effective from 3 foot candles up for normal eyes in reading legible print. In another study, Tinker (**26**) showed that subjects tend to prefer for reading approximately the illumination intensity to which they have been adapted, whether it be 8 or 52 foot candles. These data indicate that readers tend to consider comfortable for easy reading any one of a wide range of illumination intensities provided such intensities are above critical levels and provided visual adaptation is adequate. Codes of lighting have consistently ignored the role of visual adaptation in seeing. They carefully point out that the eye has evolved under the bright illumination of daylight, but do not mention that the eye also evolved to see adequately at low as well as at high intensities of light.

Illumination for Adequate Seeing

Critical levels of illumination. The critical level of illumination is the intensity beyond which there is no further increase in efficiency of performance as the foot candles become greater. Tinker (**24**) has summarized the data for critical levels of illumination: for reading of legible print (about 10 point on good paper) by adults, it is approximately 3 to 4 foot candles; for reading and study of children, 4 to 6 foot candles; for arithmetical computations, less than 9.6 foot candles; for sorting mail, 8 to 10 foot candles; for the exacting task of setting six-point type by hand, 20–22 foot candles; and for very fine discrimination required to thread a needle, 30 foot candles. In a later study, Tinker (**27**) found the critical level of illumination for reading newspaper print to be about 7 foot candles. Employing intensities from 2 to 55 foot candles, Rose and Rostas (**20**) found that reading efficiency, in terms of speed and comprehension, did not increase by a measurable amount with increased intensity of illumination.

Adequate levels of illumination. It is obvious that visual work should not be done at critical levels of illumination. There should be an adequate margin of safety to provide for individual variation and the like. For such visual tasks as reading good-sized print (10 to 11 point) on a good quality paper, i.e., print of good legibility, 10 to 15 foot candles should provide hygienic conditions when one's eyes are normal. For situations comparable to the reading of newsprint, 15 to 20 foot candles should be adequate. In situations involving the reading of handwriting and other comparable tasks, 20 to 30 foot candles seem desirable. For

tasks comparable to discrimination of 6 point type, there should be 30 to 40 foot candles. And for the most severe tasks encountered in workday situations, 40 to 50 foot candles will be found adequate. There is no valid experimental evidence now available that indicates a need for over 50 foot candles intensity for adequate visual discrimination. The intensity values from 10 to 20 should be increased somewhat (5 to 10 foot candles) for eyes with slight disabilities or for those with corrections. For the higher values, however, no practical gain will be achieved for these people by increasing the intensity. The above suggestions hold for school children as well as for adults. In general, the child has much less severe visual tasks than adults.

Intensity of illumination cannot be prescribed without coordinating it with other factors such as distribution of light and brightness contrast. A good example of the uselessness of excessively bright light is found in the study by Darley and Ickis (4). They were concerned with vision in the drafting room, a very severe visual task. In comparing 30 with 75 foot candles of indirect light, they found the efficiency ratings for the two to be only slightly different. When they compared 40 with 80 foot candles of direct light (troffer) under conditions of no reflected glare, they also found no significant differences in the efficiency ratings. The observations of Harrison (8) are relevant here. He points out the danger of glare with installations of 50 foot candles and above of artificial illumination.

Summary

Examination of the literature upon which lighting recommendations are based reveals that some techniques of experimentation are invalid, and that interpretations from certain other data are unwarranted. Some of the recommendations are adequate, others are not. The trend seems to be to specify as high intensities as the traffic will bear and at the same time to suggest to the consumer that if he uses still higher intensities he will improve his ease of seeing. All will agree that there should be sufficient light for adequate seeing. It is high time, however, that the consumer know what is adequate and what is surplus. As pointed out by Winslow (34), illumination should conform to real human needs. It is human health and comfort which are at stake.

In general the recommended practice concerning distribution of light, brightness contrast and color of light is satisfactory.

BIBLIOGRAPHY

1. BITTERMAN, M. E. Heart rate and frequency of blinking as indices of visual efficiency. *J. exp. Psychol.*, 1945, **35**, 279–292.
2. BITTERMAN, M. E., & SOLOWAY, E. The relation between frequency of blinking and effort expended in mental work. *J. exp. Psychol.*, 1946, **36**, 134–136.
3. BITTERMAN, M. E., & SOLOWAY, E. Frequency of blinking as a measure of visual efficiency: some methodological considerations. *Amer. J. Psychol.*, 1946, **59**, 676–681.
4. DARLEY, W. G., & ICKIS, L. S. Lighting and seeing in the drafting room, *Illuminating Engineering*, 1941, **36**, 1462–1487.
5. EAMES, T. H. Review of M. Luckiesh and F. K. Moss, *Reading as a visual task*, *Columbia Optometrist*, 1942, **16**, 7.
6. FERREE, C. E., & RAND, GERTRUDE. The effect of intensity of illumination on the near point of vision and a comparison of the effect for presbyopic and non-presbyopic eyes. *Transactions I.E.S.*, 1933, **38**, 590–611.
7. FERREE, C. E., RAND, GERTRUDE, & LEWIS, E. F. The effect of increase of intensity of light on the visual acuity of presbyopic and non-presbyopic eyes. *Transactions I.E.S.*, 1934, **29**, 296–313.
8. HARRISON, W. What is wrong with our 50 foot-candle installations? *Transactions I.E.S.*, 1937, **32**, 208–223.
9. HOFFMAN, A. C. Luckiesh and Moss on reading illumination. *J. appl. Psychol.*, 1947, **31**, 44–53.
10. LUCKIESH, M. *Light, vision and seeing.* New York: Van Nostrand, 1944.
11. LUCKIESH, M., & MOSS, F. K. *The new science of seeing.* Cleveland: General Electric Co., 1934.
12. LUCKIESH, M., & MOSS, F. K. *The science of seeing.* New York: Van Nostrand, 1937.
13. LUCKIESH, M., & MOSS, F. K. Illumination and eye health. *New Eng. J. Med.*, 1941, **224**, 1117–1119.
14. LUCKIESH, M., & MOSS, F. K. *Reading as a visual task.* New York: Van Nostrand, 1942.
15. LYTHGOE, R. J. *The measurement of visual acuity.* London: His Majesty's Stationery Office, 1932.
16. MACPHERSON, S. J. The effectiveness of lighting—its numerical assessment by methods based on blinking rate. *Transactions I.E.S.*, 1943, **38**, 520–522.
17. MCFARLAND, R. A., KNEHR, C. A., & BERENS, C. Metabolism and pulse rate as related to reading under high and low levels of illumination. *J. exp. Psychol.*, 1939, **25**, 65–75.
18. MCFARLAND, R. A., HOLWAY, A. H., & HURVICH, L. M. *Studies of visual fatigue.* Boston: Graduate School of Business Administration, Harvard Univ., 1942.
19. MCNALLY, H. J. The readability of certain type sizes and forms in sight-saving classes. Teachers College, Columbia Univ., Contrib. to Educ., No. 883. New York: Bureau of Publications, Teachers College, Columbia Univ., 1943.
20. ROSE, F. C., & ROSTAS, S. M. The effect of illumination on reading rate and comprehension of college students. *J. educ. Psychol.*, 1946, **37**, 279–292.
21. STURROCK, W. Levels of illumination. *Magazine of Light*, 1945, **14**, No. 4, 26–36.
22. TINKER, M. A. Cautions concerning illumination intensities for reading. *Amer. J. Optom.*, 1935, **12**, 43–51.
23. TINKER, M. A. The new "standards" for school lighting. *Sch. & Soc.*, 1939, **49**, 95–96.

24. TINKER, M. A. Illumination standards for effective and comfortable vision. *J. consult. Psychol.*, 1939, **3**, 11–20.
25. TINKER, M. A. The effect of illumination intensities upon speed of perception and upon fatigue in reading. *J. educ. Psychol.*, 1939, **30**, 561–571.
26. TINKER, M. A. Effect of visual adaptation upon intensity of light preferred for reading. *Amer. J. Psychol.*, 1941, **54**, 559–563.
27. TINKER, M. A. Illumination intensities for reading newspaper type. *J. educ. Psychol.*, 1943, **34**, 247–250.
28. TINKER, M. A. A reply to Dr. Luckiesh. *J. appl. Psychol.*, 1943, **27**, 469–472.
29. TINKER, M. A. Review of M. Luckiesh and F. K. Moss, *Reading as a visual task. J. appl. Psychol.*, 1943, **27**, 116–118.
30. TINKER, M. A. Illumination intensities preferred for reading with direct lighting. *Amer. J. Optom. & Arch. Amer. Acad. Optom.*, 1944, **21**, 213–219.
31. TINKER, M. A. Review of M. Luckiesh, *Light, vision and seeing. J. appl. Psychol.*, 1945, **29**, 252–253.
32. TINKER, M. A. Validity of frequency of blinking as a criterion of readability. *J. exp. Psychol.*, 1946, **36**, 453–460.
33. TINKER, M. A. Illumination standards. *Amer. J. Pub. Health*, 1946, **36**, 963–973.
34. WINSLOW, C. E. A. How many foot candles? *Amer. J. Pub. Health*, 1946, **36**, 1040–1041; also reprinted in *J. appl. Psychol.*, 1947, **31**, 140–142.
35. *American recommended practice of school lighting.* New York: Illuminating Engineering Soc. & Amer. Inst. Architects, 1938. Pp. 60.
36. *Recommended practice of office lighting.* New York: Illuminating Engineering Soc., 1942. Pp. 47.
37. *American recommended practice of industrial lighting.* New York: Illuminating Engineering Soc., 1942. Pp. 51.
38. *Recommended practice of home lighting.* New York: Illuminating Engineering Soc., 1945. Pp. 40.

ON FESTINGER'S EVALUATION OF SCALE ANALYSIS

LOUIS GUTTMAN

Department of Sociology and Anthropology, Cornell University

The theory of scale analysis had its origin some seven years ago. Since that time, especially by virtue of extensive and intensive research done in the Army, some of its further ramifications have been explored and several techniques have been devised for carrying an analysis out in practice. The power and incisiveness of this approach have been demonstrated in numerous attitude and opinion surveys made in the past several years, as well as in studies of achievement tests. A pleasing feature has been the simplicity of the techniques involved.

Most of the material, with respect to both applications and theoretical developments, is as yet unpublished. A manuscript has been prepared by Edward A. Suchman and the writer which will give the first comprehensive statement of both the theory and practice of scale analysis. This manuscript will form part of the four volumes soon to be published by the Social Science Research Council on the work of the Research Branch, Information and Education Division of the War Department. These volumes will also provide many illustrations of how scale analysis has been used for practical problems. Meanwhile, some brief statements of the principal concepts and instructions for practical procedures are available in article form to those who wish to use this approach in their own research (see the bibliography below).

On the basis of some articles which have been published and of some mimeographed progress reports, Festinger (1) has recently attempted a survey and evaluation of scale analysis. Since his survey is not based on all the information available, it is admittedly tentative and incomplete. In addition, full advantage has not been taken of the material which Festinger used as his sources; he raises a number of points which have already been answered there, and also introduces erroneous interpretations and conclusions.

It seems worthwhile to discuss at the present time some of Festinger's criticisms in order to help clarify the issues and to correct some important misapprehensions. Attention is also called to some articles that have appeared since Festinger prepared his paper, discussing various aspects of scale analysis (**10**, **11**, **13**).

Three of Festinger's points will be analyzed here: (a) criteria for scalability, (b) techniques of analysis, and (c) the use of scale analysis in practice. In the course of the discussion, some other aspects will be brought out which Festinger has not considered.

Criteria for Scalability

Reproducibility. The main purpose of scale analysis is to test the hypothesis that a universe of qualitative items can be represented by a quantitative variable. In order for the universe to be represented exactly by a quantitative variable, each item must be a perfect function of that variable, or be perfectly reproducible from it. Thus the concept of reproducibility is paramount in scale analysis.

In practice, only a sample of items is used from the universe of content. Furthermore, in practice, it is not expected to find perfectly reproducible or scalable universes. Among other things, perfect reproducibility implies perfect test-retest reliability, which is certainly not to be expected empirically. However, if the reproducibility of the entire universe is very high, say over 90%, then that may be sufficient for many practical purposes. A quantitative variable which will represent an indefinitely large universe of items that well will ordinarily not lose much predictive power, whether used for predicting outside variables or whether predicted from outside variables. This will especially be true if the errors of reproducibility are random.

Since universe reproducibility must be estimated on the basis of only a sample of items, it becomes evident that the sample's reproducibility alone may not be a sufficient guide. Festinger criticizes the sample reproducibility coefficient for its inadequacy.* This inadequacy was recognized at the outset in scale analysis. The same kind of examples that Festinger uses (1, pp. 156–157), showing how five or nine statistically independent items can have high reproducibility, were worked out previously; several such examples will appear in the forthcoming volume. Indeed, there is an even worse case than that of statistical independence, namely that wherein some items have *negative* relationships with others; this is worse than being statistically independent from the point of view of scale analysis. Examples can be constructed showing how even in this case it is possible to have supriously high reproducibility in a small sample of items.

Festinger omits to point out that this problem about reproducibility was raised before, and that several answers have already come forth. In one of my mimeographed reports to which Festinger refers (15), there is the following question and answer:

Q. *Is reproducibility by itself a sufficient test of scalability?*
A. No. It is the principal test, but there are at least three other features

* Hau Knecht (12) has raised this criticism earlier, also without taking cognizance of the fact that other criteria have always been used as discussed below.

that should be taken into account: (a) range of marginals, (b) random scatter of errors, (c) number of items in the sample.

Further questions and answers elaborate on the point. And again, in another paper (7) to which Festinger refers, it is stated:

> The percent reproducibility alone is not sufficient to lead to the conclusion that the universe of content is scalable. The frequency of responses to each separate item must also be taken into account for a very simple reason. Reproducibility can be artificially high simply because one category in each item has a very high frequency. It can be proved that the reproducibility of an item can never be less than the largest frequency of its category, regardless of whether the area is scalable or not.

And further:

> An empirical rule for judging the spuriousness of scale reproducibility has been adopted to be the following: No category should have more error in it than non-error.

If this latter rule alone were applied to Festinger's examples, it would immediately reject the hypothesis that the items are from scalable universes. The consideration about pattern of error would also disqualify the hypothesis that the items were from scalable universes.

An Alternative. One contribution to spuriously high reproducibility is the fact that each item is being related to a score which is based in part on the item. An alternative way to compute the coefficient of reproducibility is to hold out each item in turn from the analysis, thus obtaining N sets of trial scale scores. The errors for each item can then be counted from its relationship to the score based on the $N-1$ other items.

If this *partial-score* method were used on *statistically independent* items, then the reproducibility for each item would be precisely the relative frequency of its modal category. Thus, in Festinger's example (1, p. 156) of five independent dichotomies with marginals 80%, 60%, 50%, 40%, and 20% the respective modal relative frequencies are 80%, 60%, 50%, 60%, and 80%; hence, the reproducibility of all five items, computed from partial scores, would be the mean of the latter five percentages or 66%, compared with the spurious 86% Festinger obtained from whole scores. Indeed—*no matter what the interrelations of the five items were*—their reproducibility could not be less than 66%, because reproducibility of an item can never be less than its modal frequency. Similarly, in Festinger's second example (1, p. 157) of nine statistically independent dichotomies with marginals .9, .8, .7, .6, .5, .4, .3, .2, .1, the respective modal proportions of the items are .9, .8, .7, .6, .5, .6, .7, .8,

and .9, so the reproducibility of the set cannot be less than .72; Festinger finds .83 reproducibility from whole scores whereas if part scores were used the obtained reproducibility would be .72.

Items with extreme marginals like .9 and .1 do not help much in testing reproducibility since such items can never have more than 10% error.

In practice, it does not usually seem worthwhile to bother with partial scores, although this technique is available for doubtful cases. The fictitious examples of independent items do not illustrate what is to be expected in practice. Attitude (or achievement) items of the same general content are usually sufficiently correlated so that scores based on eleven of them will not be substantially different from scores based on twelve. Reproducibility from whole scores will not be much greater than from part scores—so their spurious excess of reproducibility over that from part scores can be largely ignored. Furthermore, even part-score reproducibility is not a sufficient test of scalability, for the additional criteria mentioned above must also be considered.

There is room for more improvement on criteria for scalability when samples of content are used, but it should be made clear that reproducibility by itself has not and is not the sole basis for drawing inferences from a sample of items. It is the basic one, because the reproducibility of the universe is essentially what is in question, but additional criteria have been and are being used.

Reliability. The suggestion that Festinger makes that the expected occurrence of scale responses be calculated under the assumption of a perfect scale plus a certain degree of unreliability is a promising one. This idea had been thought of in the earlier stages of the development of scale analysis but discarded in the form Festinger has suggested. The proportion of people with no scale errors *cannot* be properly calculated by the method that Festinger uses. Apparently he assumes that if .9 is the proportion of population responses that will be in the scale pattern for one question, then the proportion that will be jointly *within* the scale pattern for seven questions is $(.9)^7$ or 47.8%. Unfortunately, the same reasoning would say that the proportion of people who have seven responses *outside* the scale pattern should be $(.1)^7$; and in general the proportion of people with X scale responses and $7-X$ scale errors should be given by the binomial distribution

$$\frac{7!}{X!(7-X)!}(.9)^X(.1)^{7-X}.$$

But this is impossible, for nobody can have *all* his responses as scale

errors. Indeed, for the empirical example that Festinger borrowed (1, Fig. 2, p. 157), no matter what pattern of response a person may have, he can be placed into one of the scale patterns with at most four errors. Therefore, the range of possible errors for each person is 0 through 4, rather than 0 through 7 as Festinger supposes. This means that Festinger's calculations cannot be carried out consistently to estimate reproducibility under the given assumption. The difficulty is that whether a person will fall into the scale pattern is *not* independent of whether another of his responses is within the scale pattern. Unreliability does not behave that way with respect to the scale pattern.

The actual reproducibility of this example of seven questions was about .85 rather than the .9 Festinger assumed. It is interesting to note that $(.85)^7$ is .32, which is not far from the "over one-fourth" perfect scale types reported. Actually, the universe sampled by these seven questions would not now be accepted as sufficiently scalable but would be broken up into sub-universes; the study was made when 85% reproducibility was the empirical rule rather than the present 90%. The study did serve its purpose well, however, as collateral evidence presented there showed.

The further calculation that Festinger makes of adding 3.7% to his 47.8% seems based on an unfortunate double usage of the word "chance." In his second paragraph on p. 158 (1), "chance" is used to mean statistical independence between items. Such independence cannot exist simultaneously with the assumption of a scale pattern in his following paragraph; that is, the 7% who fall into perfect types under the hypothesis of independence of items have nothing to do with the distribution of error under the assumption of uni-dimensionality plus unreliability. The binomial distribution by itself—if it were correct—takes care of the second situation. Hence, Festinger's calculations are incompatible in adding 3.7% (7% of $1-.478$) to 47.8% to obtain 52.2% as the "chance" proportion. The 7% is correct for independent items; the 47.8% would be correct for the scale-plus unreliability case if the binomial hypothesis held; and the two cases do not hold simultaneously. "Chance" means something different in each case.

A consistent use of reliability. Several correct approaches to the use of the concept of unreliability are possible, instead of the inconsistent binomial approach. One such approach will be sketched here briefly for the case of dichotomies. Let n be the number of dichotomies in the sample of items so that there are $n+1$ scale types or ranks possible. Let r be the rank of the type that is "positive" on r of the items; r ranges from 0 to n. Let Pr be the proportion of the population whose

"true" rank on the n items is r, and let Prj be the probability a person of "true" rank r will be "positive" in the jth item ($j=1,2,\ldots,n$). There are 2^n types of people—scale and non-scale—possible on the n dichotomies. The expected proportion in each of the 2^n types can be calculated from the Pr and Prj. Conversely, from the observed 2^n proportions in an actual experiment, the Pr and Prj can be estimated. There are $n+1$ parameters Pr, of which n are independent since their sum must be unity. There are $(n+1)n$ parameters Prj, all of which are independent. Hence, there are $n+(n+1)n$, or $n(n+2)$, independent parameters to be estimated from 2^n-1 independent observations. If n is greater than 5, this provides more equations than there are unknowns—*so the hypothesis of the scale structure can be tested*, as well as having the parameters estimated. Unfortunately, the equations involved in the above analysis are curvilinear, and do not seem to lend themselves to practical use because of the difficulties in the numerical computations. Furthermore, even this analysis has been simplified by assuming that persons within the same "true" rank were equally reliable within each item. Without this simplifying assumption, the equations would have innumerably more parameters.

In any analysis using the concept of test-retest reliability, it must be remembered that scalable data must in general be highly reliable, although the converse is not necessarily true. The coefficient of reproducibility—especially if computed by the part-score technique described above—sets a *lower bound* to the average reliabilities of the items (6, and especially 8). In particular, if items are perfectly reproducible they are perfectly reliable. Hence, Festinger errs in his assertion that "Even if a perfect scale were achieved these claims [concerning invariance properties] would all be limited by the degree of reliability . . . of the questions asked" (1, p. 160). Perfectly scalable data are perforce perfectly reliable. Conversely, highly unreliable data cannot be scalable. One of the contributions of a scale analysis is to provide automatically information about reliability by helping set a lower bound to it for each item.

The simple criteria used in conjunction with that of reproducibility for sample data do serve to distinguish between data that are highly scalable and those that are not. The case where the items are independent will always be rejected on the basis merely of the criterion of improvement, namely, that no category should have more errors than non-errors. The further criteria of studying patterns of error also tend to insure that no dominant second variable is present even if reproducibility is high. That is what is meant by the statement that "in imperfect

scales, scale analysis picks out deviants or non-scale types for case studies." If no non-scale types have substantial frequencies, then that tends to indicate that there is no substantial second factor present. However, if one or more non-scale types do have a substantial frequency, then that is an indication of where an additional factor (or factors) is entering into the picture. If an additional factor is sufficiently prominent, it may be worthwhile to try to piece it out further by asking additional questions. The universe might be divided into two or more subuniverses, each of which may be scalable separately. Or it may turn out that the additional factor is so highly correlated with the most dominant factor that it does not make much difference whether they are treated as two separate variables or as a single variable.

The problem is not to find out whether a perfect scale is present in practice, but rather whether it is worth worrying about any additional variables that may be present. The criteria used in practice are believed to provide an answer to this and to decide properly whether or not a set of data can be regarded as sufficiently scalable for most practical purposes.

Quasi-scales. One kind of non-scalable universe is called a *quasi-scale.* A quasi-scale is different from a scale, not just in the reproducibility, but in *the entire pattern of responses.* Festinger seems to have misunderstood the definition of a quasi-scale, for he seems to believe that it differs from a scale only with respect to reproducibility (1, p. 156 and p. 159). A universe which is quasi-scalable will ordinarily have less than 85% or so reproducibility, but that is not its distinguishing feature. The distinguishing feature is *the gradient in the responses to the items.* Cutting points cannot be established (as in the case of a scale) which will enable one to say that a person above the point is in one category of an item and a person below the point is in another category; but one can state that, if one person is higher than another in the quasi-scale, then his probability of being in a higher category of an item is correspondingly greater.

There are many kinds of configuration which are less than 85% or 90% reproducible and which are not quasi-scales at all. For example, an area may have two or more dominant factors in it, in which case it would not be either a scale or a quasi-scale. In a quasi-scale, there are one dominant factor and infinitely many small factors. The order of people in a quasi-scale is according to the dominant factor, and is essentially invariant from sample of items to sample of items, provided that the samples are large enough. There is a great deal of work yet to be done on the theory of quasi-scales, but enough is known to say that

they have quite a different character from scales and from other kinds of universes. Another distinguishing feature between a scale and a quasi-scale is that the scale has an intensity function and further meaningful components, whereas a quasi-scale does not have an intensity function or further components of that kind.

Neurotic phenomena have been found to be quasi-scalable. For example, the Neuro-Psychiatric Screening Adjunct, which is the official paper and pencil test used at all military stations since October, 1944, is a quasi-scale and is a product of a rigorous investigation of efficient screening tests made possible by the scale analysis approach (16, 17).

Techniques for Scale Analysis

Scalogram devices. There are several alternative schemes now available by which to carry out a scale analysis in practice. They are virtually equivalent in terms of the results they yield, but they differ somewhat in operation. *Scalogram boards* have been the principal device used by the War Department, and are perhaps the most flexible and easiest to use. The boards are relatively simple to make and to operate; the cost depends upon how large a board is desired and whether or not a pair is to be made. If a single board is used instead of two, then the workmanship need not be precise and the board can be made fairly cheaply by any carpenter. There are alternative mechanical schemes that might be used instead of the wooden board, and undoubtedly other schemes will be invented in the future which will be even easier to construct. Instructions for the construction and use of a scalogram board will appear in the forthcoming volumes on the work of the Research Branch.

The Cornell technique (7) is also very easy to learn; it is taught in a course on attitude and public opinion analysis to students who have no background whatsoever in statistics. For achievement tests, where all items are dichotomous—being marked either right or wrong—the Cornell technique is perhaps the best of all to be used. For dichotomies, there is no problem of combination of categories, so that there is but one trial to be made in an analysis. The Cornell technique suffers a bit in flexibility compared to the scalogram board when a series of trials has to be made. Ordinarily, but two trials may be needed in an analysis, and the Cornell technique has proved very advantageous in such cases for general research purposes. It can be carried out on IBM equipment as well as by hand.

The Goodenough technique (2) is based upon an explicit tabulation of all combinations of responses that actually occur. It is more "rigor-

ous" than the preceding two techniques in that it counts the errors at each stage. However, it yields no different results in the end. Apparently Festinger has not worked through the Goodenough technique to see how it does work out in practice.* The first step seems simple, but it takes a good deal of experience to master the three following steps. The process becomes very bulky and involved when ten or twelve items are used.

The Cornell Technique has the advantage that its complexity does not at all change, regardless of the number of items (though of course the amount of labor increases with the number of items). The same lack of increase of complexity holds to a slightly less degree with the scalogram board.

The problem of metric. The earliest technique for scale analysis was that of least squares (3). It is quite properly to be abandoned as a procedure in practice because it is certainly far more cumbersome than the others. However, the equations involved have turned out to be of basic importance in *interpreting* a scale, and have led in particular to the empirical treatment of the intensity function which is proving so vital for attitude and public opinion work. Also, the basic thinking behind the equations have led to a solution to the related problem of paired-comparisons (12).

In the beginning of my work on scale analysis, I had thought that one of the most important problems was that of metric. I had thought that how to obtain weights for items was perhaps the leading problem to be solved. But as the theory of scale analysis developed, it became clear that the problem of weights was essentially a minor one for most practical purposes. Indeed, for the *perfect* scale pattern, it is easy to see that if scores are to be obtained for people by adding up weights assigned to categories of items, then, no matter what weights are used—as long as they have the proper rank order within each item—the scores of the people will have *exactly* the same rank order. The ordering of people in this sense does not depend at all upon finding a particular weighting system.

The important problem turned out to be that of finding the *structure* a universe of items must have in order to be scalable; it was not that of finding weights.

* Festinger also apparently has misread Goodenough as to how to measure reproductibility. Goodenough explicitly says that "at least 85% of the total number of responses must fall within the scale pattern, so that it is possible to reproduce 85% correctly all the responses of all the respondents from the scale scores" (2, p. 184). Festinger seems to have misread this to mean that 85% of the individuals fall into perfect scale types.

The problem of a metric does turn out to come into the picture for further problems, and it first appeared as a practical problem with respect to that of bias in questionnaire wording (**11, 13, 14**). The problem here was, after people are ranked from a high to a low on an attitude or opinion, to find a dividing point in the order such that the people on one side can be called positive and people on the other side can be called negative. The equations of scale analysis, when applied to the perfect scale pattern, show a most remarkable result. They show that a universe of items which is perfectly scalable can be resolved into an infinite series of principal components, the first of which provides the basic metric, the second of which is the intensity component, and the remaining ones are as yet not named (**10**). Empirical study of the intensity function has afforded for the first time a scientific solution to the problem of question bias.

These equations, then, show that a scalable attitude is somewhat different from the twelve-inch ruler that Festinger uses as an analogy (**1**, p. 160). The responses of a person to items in a scalable universe are seen by means of these equations to be a function of the person's metric score, his intensity, and the further components in the scale. The person's rank order is sufficient to reproduce his responses exactly; in this sense, the responses of the population are but a function of a single variable. Resolving the responses into components by the alternative device of the least squares equations shows the responses to be a function of infinitely many variables, each of which is a function of the rank order.

These striking results from using the least squares equations in conjunction with the perfect scale pattern will be elaborated on in the forthcoming publication on the work of the Research Branch. It might further be pointed out here that these equations resolve also the paradox which appears in achievement tests where the *difficulty* of an item seems to introduce a factor different from the common *content* factor that the items may have. Since scale analysis applies to achievement tests as well as to attitude or opinion areas, achievement tests also are resolvable into the principal components of a scale. In a scalable achievement test, then, each item is a function of but a single dimension from the point of view of reproducibility, but a function of infinitely many dimensions from the point of view of principal components. The apparent contradiction between these two points of view is resolved by the fact that the infinitely many principal components in turn are perfect functions of the rank order of people.

Uses of Scale Analysis

Incidence of scales. The theory and techniques of scale analysis provide a test of the hypothesis that a universe of qualitative items can be represented by a single quantitative variable. This hypothesis is appropriate for any qualitative universe obtained by any method of observation. The universe may be a set of items recorded on a questionnaire, or observations obtained in non-directive interviews, by participant observation, or by any other technique of gathering data. No matter how the data are gathered, each observation is but a sample of all similar observations that could have been obtained, and the entire universe of observations is ordinarily of interest.

As Festinger suggests, scalable universes may be the exception rather than the rule. Festinger does not give any explicit reasons for his belief, but this position will be substantiated in the forthcoming volume. It has already been pointed out that one possible reason for the existence of an attitude scale is that of a homogeneous culture (**4**, p. 149). If a population is not subjected to the same social stimuli with respect to the attitude, it might be expected that it will prove to be unscalable for them. The fact that neurotic phenomena have not been found scalable can perhaps be explained in this fashion. Similarly, an area of achievement may be expected not to be scalable if there is no uniform program of training for the population involved.

Another reason for expecting many universes not to be scalable in practice is that the notion of a universe is so comprehensive. Each sub-universe of a universe is of course itself a universe. Since there is ordinarily a vast number of imperfectly related sub-universes, there must be a vast number of combinations of them which are non-scalable universes. Merely this formal consideration would lead one to believe that most universes are not scalable. Non-scalable universes may of course be broken down in some cases into scalable sub-universes. One of the contributions of scale analysis is to point out the need for being clear about the universe's content. By focusing on more and more homogeneous content, research can be made more meaningful and external predictions be made more effective in the long run.

The development of the above-mentioned screening test for psychoneurotics (**16**, **17**) is but one example of how research utilizing scale analysis was more effective than it would have been had the more traditional but less incisive procedures been followed. Instead of throwing together all kinds of conceivable predictive items into one composite, fifteen different universes of content were defined which might be re-

lated to the criterion of psychoneuroticism. The structure of each of these universes was first analyzed separately. Because each was found to be either a scale or a quasi-scale, only a relatively few items from each were needed in order fully to utilize the predictive power of the universes. The multiple correlation of the criterion was then worked out on all fifteen predictors with the finding that one of the universes predicted as well as the best combination of the fifteen. This enabled the short but efficient screening test to be used with the knowledge that it retained the predictive power of innumerably many items in fifteen different universes. Such a complete usage of predictive power could not have been made without scale analysis.

From the practical point of view, another important feature here is the amount of labor saved by scalogram techniques in obtaining this maximum predictive power, compared to using more traditional techniques which are far more laborious and which would yield less effective predictions.

The two problems, that of scalability and that of external prediction are distinct but related. By focusing on the scaling problem in its own right, more effective external predictions are thereby made possible.

There are many areas which have been found to be scalable thus far, and therefore these areas can be handled economically by means of simple scale scores. Many areas have also been found not to be scalable; all such areas cannot be handled so simply. It is known how to treat quasi-scalable areas, and Lazarsfeld is now completing a theory of the latent dichotomy which also can be handled by means of a single quantification. How to utilize other kinds of non-scalable areas is still an unsolved problem. The emphasis that scale analysis makes in this connection is that unless the structure of the universe is known, it is not known how best to treat the universe for any particular purpose.

Distinction between theory and techniques. The basic *theory* of scale analysis is not to be confused with particular *techniques* for carrying out such an analysis in various kinds of situations. Festinger borders on confusing the two when he states that " 'scale analysis' seems to be an excellent technique for use with paper and pencil tests or other instances of measurement where the situation permits the inclusion of several questions centering about the same topic" (1, p. 160). If a research problem is concerned with a universe of content, then that universe must be studied. That is what the theory calls for. How adequate is the technique which Festinger implicitly advocates of studying only a single item from the universe?

One of the important aspects of a universe of content is its structure;

for example, is the universe scalable or does it have some other kind of structure? The theory of scale analysis tells what a scalable structure is, and the various properties possessed by such a structure.

The practical problem is to obtain information about the structure from only a sample of items. It has already been indicated how an adequate sample of items can be chosen to test the hypothesis of scalability. Furthermore, the number of items to be used in a pretest must be distinguished from the number of items to be used in a final study. One of the properties of a scalable universe is that only one or two items can be used in a final study for many purposes once their place in the universe is ascertained. The scalability of the universe must first be analyzed, however, by a dozen or so items in a pretest.

The statement that "most of those engaged in this type of research [public opinion] will probably find the inclusion of a series of questions which could be subjected to scale analysis not feasible from practical considerations" (1, p. 159) does not accord with what is the actual practice both in public opinion and in market research, as well as in general attitude research. It is because workers in these fields are concerned with a universe of content that they pretest various questions on the same topic; it is a foolhardy pollster who bases conclusions on but a single question. The use of the split-ballot is evidence of this concern with sampling of content. In addition, ordinary polls often include several questions on the same topic on the same ballot. The extreme position taken by advocates of "open-ended interviewing" is to ask a whole series of questions of every respondent. And of course, conventional attitude surveys almost invariably use a substantial set of questions for a given topic.

It is a misapprehension to believe that asking several questions on the same topic necessarily creates a problem of rapport. In one survey made of a national cross-section by a leading public opinion polling agency, an area of content was defined and then sampled by four questions. Some of the interviewers complained because of the great similarity of wording of questions. The questions were very similarly worded because the content concerned the size of the Navy and was very hard to discuss in different ways. But even under these adverse circumstances, the analysis was successful in showing that the area was scalable and that the zero point could be located properly by the intensity function. Even more questions in the same area had been used in the pretest in Ithaca on a cross-section of the population there, and interestingly enough there was no complaint either from the respondents or from the interviewers, although the interviewers were no different from those

used in the national cross-section and had no knowledge whatsoever of what was involved in scale analysis. An area of apparently very similar questions is an exception rather than the rule. The example about desire for post-war schooling that Festinger has borrowed (1, p. 157) certainly provides no problem of rapport, and the general run of areas studied by public opinion polls do not present any particular problem of rapport. Another large market research agency has tried scale analysis in a routine study and has found no difficulty whatsoever with it. Because of its simplicity and its objective solution to the problem of bias, this agency plans to use this approach regularly.

It seems premature, then, to conclude that scale analysis cannot be carried out in practice in public opinion work. To the contrary, scale analysis is becoming more essential in this field because it affords for the first time a scientific solution to the basic problem of bias in public opinion polls. This problem arises from the fact that a universe of content is being studied and any single question is but a sample of all possible questions that could have been asked. How can one determine which question does coincide with the zero point of the entire universe, that is, the point which divides those who are negative on the issue from those who are positive?

The intensity function provides a scientific solution to this problem (13). It provides both a definition and a technique for ascertaining a zero point for the population. Unless some such objective approach to the question of bias is used in public opinion polls, it cannot be certain how much credence to place on their reports.

By providing a solution to the problem of bias, scale analysis clears the way for asking questions in the manner which will best help establish rapport with the respondent. The particular form of a question does not affect the results of scale analysis, so the research worker can concentrate on obtaining the wording which will make the interviewing work go most smoothly. Thus scale analysis has a contribution to make toward increasing rapport in surveys rather than the contrary. Apprehension that the opposite is true seems to be due to a misconception that scale analysis presupposes a particular way of asking questions.

If progress is to be made in the scientific study of attitudes, public opinion, and achievement, it seems necessary to concentrate on the problem of the structure of content. Techniques are not worth much if not guided by any theory. The theory of scale analysis happens to lend to simple and practical techniques. To compare these techniques with others, one would have to ask: what theory of structure guides the alternative techniques and how adequately is this theory served thereby?

BIBLIOGRAPHY

1. FESTINGER, L. The treatment of qualitative data by "scale analysis." *Psychol. Bull.*, 1947, **44**, 149–161.
2. GOODENOUGH, W. H. A technique for scale analysis. *Educ. psychol. Msmt.*, 1944, **4**, 179–190.
3. GUTTMAN, L. The quantification of a class of attributes: A theory and method of scale construction. In P. Horst *et al.*, *The prediction of personal adjustment*, Soc. Sci. Res. Council Bull. No. 48, 1941, 319–348.
4. GUTTMAN, L. A basis for scaling qualitative data. *Amer. sociol. Rev.*, 1944, **9**, 139–150.
5. GUTTMAN, L. Scale and intensity analysis for attitude, opinion, and achievement. (Mimeographed, 1945. To appear in the proceedings of the conference on military contributions to methodology in applied psychology to be published by the Univ. of Maryland Press.)
6. GUTTMAN, L. A basis for analyzing test-retest reliability. *Psychometrika*, 1945, **10**, 255–282.
7. GUTTMAN, L. The Cornell technique for scale and intensity analysis. (Mimeographed, 1946. To appear in the proceedings of the conference on measurement of consumer interest to be published by the Univ. of Pennsylvania Press; also in *Educ. psychol Msmt.*)
8. GUTTMAN, L. The test-retest reliability of qualitative data. *Psychometrika*, 1946, **11**, 81–95.
9. GUTTMAN, L. An approach for quantifying paired comparisons and rank order. *Ann. math. Statist.*, 1946, **17**, 144–163.
10. GUTTMAN, L. Suggestions for further research in scale and intensity analysis of attitudes and opinions. Int. J. Opin. & Att. Res., 1947, **1**, 30–35.
11. GUTTMAN, L., & SUCHMAN, E. A. Intensity and a zero point for attitude analysis. *Amer. sociol. Rev.*, 1947, **12**, 57–67.
12. HAUSKNECHT, G. A procedure for determining a useful approximation to an ideal scale. Unpublished manuscript.
13. SUCHMAN, E. A., & GUTTMAN, L. A solution to the problem of question bias. *Publ. Opin. Quart.*, 1947. (In Press.)
14. *Experiments on the measurement of the intensity function and zero point in attitude analysis.* Research Branch, Information and Education Division, Army Service Forces. Report D-1, Mimeographed, 1945.
15. *Questions and answers about scale analysis.* Research Branch, Information and Education Division, Army Service Forces. Report D-2, Mimeographed, 1945.
16. *The screening of neurotics.* Research Branch, Information and Education Division, Army Service Forces, Report B-104, 1944.
17. *A study of psychoneurotics in the army.* Research Branch, Information and Education Division, Army Service Forces. Report B-107, 1944.

NOTE ON "A REVIEW OF LEADERSHIP STUDIES WITH PARTICULAR REFERENCE TO MILITARY PROBLEMS"[1]

DONALD E. BAIER

Personnel Research Section, A.G.O.

The valuable report[2] with which this note is concerned ". . . summarizes and reviews selected references from the available literature dealing with the problem of the selection of leaders in various fields. The primary interest in preparing the article was to provide a summary of techniques and results that would be of value to psychologists dealing with problems of selecting leaders, particularly in the military field."

It is the purpose of this note to make available additional facts and comments which appear to bear on the following conclusions of the reviewer:

1. "Progress has not been made in the development of criteria of leadership behavior . . . ."
2. "Advances in methodology in this field are definitely not striking."

It is this writer's belief that these conclusions, insofar as they are meant to apply to military leadership, are not entirely warranted.

In two reports[3] published by the Medical Field Research Laboratory, Camp Le Jeune, N. C., research on measurement of "leadership" is reported. These studies indicate a substantial relationship (tetrachoric $r = .42$) between superior officers' reports of the combat performance of Marine Corps officers graduated from the Corps Officer Candidate School and the standing of these graduates among their fellow-marines as indicated by a nomination procedure conducted during their pre-officer training. The two sets of evaluations were completely independent.

An as yet unpublished follow-up study by the Personnel Research Section, AGO, of West Point graduates after 18 months of duty as Army officers also reveals a significant association ($r = .51$ for Infantry Officers) between inter-cadet ratings or leader-nominations and success as an officer measured by the *Officer Efficiency Report, WD, AGO Form 67*. Here again there is basis for believing that the two measures are independent.

[1] The opinions expressed herein are those of the author and do not necessarily represent the official view of the War Department.

[2] JENKINS, WILLIAM O. A review of leadership studies with particular reference to military problems. *Psychol. Bull.*, 1947, **44**, 54–79.

[3] *Validation of officer selection tests by means of combat proficiency ratings.* Medical Field Research Laboratory Report No. 1, January 18, 1946 and No. 2, May 16, 1946. Camp Le Jeune,

The reviewer's account of the research upon which are based the current methods for selecting wartime officers for integration into the regular Army may result in misunderstanding. In discussing the correlation between the Officer Evaluation Report and the criterion of leadership, the latter being a product of nominations by subordinates and peers with a veto power resting with the commanding officer of the group, Jenkins states:

> . . . The degree to which the Commanding Officers' ratings were weighted in the Officer Evaluation Report was not stated, but it appears likely that this factor played an important role. Substantial agreement between ratings by the C.O. and by fellow officers was to be expected. Since the OER had the highest validity, and the other measures when combined with it increased its correlation with the criterion only .07, these questions suggest the necessity for a further examination of the nature of the criterion here employed (p. 74).

The Officer Evaluation Report was accomplished in the majority of cases by the immediate supervisor, not the C.O., and represented only the former's evaluation of the ratee. The conclusion that substantial agreement between ratings by the C.O. and by fellow officers was to be expected does not appear to be justified. The C.O. was only one of from 7 to 30 nominators who participated in determining the ratee's criterion standing. He had no knowledge of how the other members of the nomination group evaluated each ratee, and his rating was used only to eliminate from the criterion groups of High, Low or Middle those rare cases where the C.O. placed the rated officer in the opposite extreme from the combined ratings of his subordinates and peers. Later studies employing a criterion which did not include the C.O. showed no drop in the validity of the OER or FCL type rating device. It is our belief that the nomination criterion employed in the studies cited does represent progress in the development of leadership criteria.

With respect to methodology, the forced choice technique as exemplified in the triads and tetrads of the OER and the recently revised Army Officer Efficiency Report seems to deserve more attention than the reviewer accords it. This technique, which has been described briefly in a paper titled "The Forced Choice Technique and Rating Scales," presented at the American Psychological Association meeting in Philadelphia on 5 Sept. 1946 by the Personnel Research Section, AGO, not only provides valid indicators of the ratees' standing on a nomination criterion, but favorably influences ratings of overall competence (if they are made immediately following completion of the FCL items) so that they show substantially less negative skewness. Clearly, the forced-choice technique is effective in diminishing rater-bias and in improving the distribution and validity of ratings which are generally regarded as indicative of leadership performance.

BOOK REVIEWS

MUNN, NORMAN L. *Psychology: The fundamentals of human adjustment.* Boston: Houghton Mifflin, 1946. Pp. xviii+497.

The importance of the introductory course in psychology cannot be overestimated for it determines to a great extent the student's attitude toward the subject and whether or not he goes any further with it. But the importance of the textbook depends to a large degree upon the instructor. Some instructors lean heavily upon the text, others hardly at all. In reviewing a book, however, evaluation must be made as if it were the sole source of the student's introduction to the subject, regardless of the instructor's predilections, interpretations, choice of material, or method of handling the course. Though there are suggested readings at the end of each chapter in this as in other texts which the student is urged to consult, their influence is admittedly minor since the author of a text, as Munn says, writes with the feeling that he, in common with most teachers of the subject, could "organize its topics in a more logical sequence, choose apter illustrations, find more interesting examples and . . . write a book that . . . would be more appealing to instructors and students than any he has seen" (ii). Some other requirements of a good introductory text are succinctly suggested by President Leonard Carmichael in his Introduction wherein he discusses the reasons for studying Psychology today: as an essential part of a general education; as preparation for professions like law, medicine, teaching, the ministry, and business; and for further professional work in the subject itself. That Munn has met both his own and the editor's demands with considerable success there can be no question. The book is plentifully provided with excellent diagrams, half-tones, and tabular matter; it is full of concrete material chosen from a wide variety of sources; and its approach is scientific throughout.

It would seem, in the light of these virtues, that this text meets the requirements of the introductory course almost to perfection. It is such a splendid job in so many respects that any criticism at all seems supererogatory, if not hypercritical; and yet there are qualities expected in an elementary text which are of equal or greater importance than the ones which this book has in such large measure. The most important lack is an underlying point of view or theoretical structure integrating and unifying the topics and their relations to each other and the subject as a whole. We shall find that in this respect the book is not up to its accomplishments otherwise.

The book is divided into seven main parts and these in turn into two or more chapters. It begins with a discussion of general, methodological and historical material, then proceeds to consider in turn the anatomical and physiological bases of behavior, learning, remembering, thinking,

motivation, conflict, feeling, perceiving, the special senses, statistics, intelligence, aptitudes, and personality. The treatment develops by consideration of simpler processes followed by the more complex, in so far as possible, though there is some back-tracking which is done with a minimum of repetition of earlier material. With the general plan of the work before us we can now consider it more in detail.

Beginning with the origin and scope of psychology, the first two chapters are devoted to a brief glance at the history of the subject through a consideration of such topics as the psyche, the organism, methods in philosophy, physiology, and physics, analysis of consciousness and some fields of psychology. Chapters 1 and 2 really constitute a single topic or set of topics and furnish an excellent survey of methods, fields, and problems. They are properly brief, to the point, and very readable. Only in one detail does the text here need emendation. In discussing scientific controls it is stated that "there is never more than one independent variable in a given experiment . . . . If two or more factors were varied, he (the scientist) obviously would not know which had produced the phenomena observed" (p. 23). While it is not expected that the logic of analysis of variance and designed experiments should be presented at the elementary level (though it is not impossible) advances in statistics have made the older Mill-Bacon canons of scientific procedure represented by this statement quite out-of-date. Variation of only a single variable in psychological experiments is possible so seldom as to be almost a fiction and now that we have the statistical tools for handling multiple variates we might as well give up the fiction.

Part 2 deals with psychological development and consists of three chapters: origin and psychological significance of response systems, conception to maturity, and factors in psychological growth. Here the biological bases of behavior are explained and the psychological processes most directly correlated with them are brought in. The result is that the simpler and more complex processes are more or less intermingled in these chapters as a partial list of the topics reveals: structure and functions of receptor and nervous systems, embryonic development, sensitivity, locomotion, prehension, language, gestures, writing, genes, heredity, environment, and maturation. The order is in general from simple to complex but there are some reversals. Thus one would expect a discussion of genes and embryological development before discussion of the nervous system but here it follows the latter. The reason for Munn's order is obvious to a reader of the book and a good one: discussion of the more elementary gene units links directly with problems of heredity, environment, maturation and growth. There is no discussion of nerve action potentials and the treatment of the autonomic nervous system is postponed to the chapter on emotion where diagrams illustrating its relations to the cerebro-spinal system are given. The reviewer finds it impossible to omit the autonomic nervous system when

explaining the rest of the nervous system, although like Munn, he finds greatest use for it in the discussion of feelings and emotions. Figure 13 in chapter 3, showing the spinal reflex-arc system, would not be overcomplicated if the sympathetic ganglion and its fibers were included as is usually done, and with some textual discussion this would remedy a serious omission at this point. On the other hand, Munn has included more material on the nervous system than is generally presented. The diagrams showing different types of synaptic connections make interaction, facilitation, and inhibition intelligible neurologically. The discussion of cortical representation of sensory functions is especially well done.

The chapters on conditioning, learning, memory and thinking succeed in presenting a considerable amount of material but suffer from the lack of clear integrating principles. While Munn rejects classical, Pavlovian conditioning theory as an adequate account of all learned behavior it is not clear what principles he would employ instead. That the author places greatest reliance on trial and error, past experience, and association appears from his treatment of certain particular problems rather than from explicit structuration of the material. One must dig his underlying approach out from a few critical cases which reveal the author's fundamental position. Thus the explanation of how the chimpanzee reaches an apparently inaccessible object is a case in point. According to Munn we are to suppose " . . . a chimpanzee has, in the jungle, learned to reach otherwise inaccessible objects by swinging toward them on a vine. Now in the psychological laboratory, he is confronted with an apparently inaccessible banana. A rope, however, is hanging nearby. *If the animal sees the similarity between the rope and the vine, or between his jungle method and the one now possible, he may solve the problem immediately*" (p. 122). (Italics are the reviewer's.) Now this explanation of the ape's accomplishment in terms of past experience which at first sight seems to be the scientifically simplest explanation actually turns out on closer scrutiny to demand much more in the way of memory and intellectual ability than the proposition that the animal simply sees the relevance of the rope to the banana which is immediately given. This assumption should not be too difficult to make since it was pointed out earlier that the difference between classical (mechanical) and instrumental conditioning lies in the fact that conditioning occurs much more easily when in the direction of more relevant responses. Certainly if the principle of relevance is basic to conditioning it may be accepted for the much more complicated case of insightful behavior. The welter of factors involved in acquiring skill and learning could be better ordered and made more meaningfully connected if some structure were seen behind the facts in question.

The lack of an adequate theoretical framework plagues the reader most in the concluding chapter on thinking. Reasoning, we are told, is

implicit trial and error, it is a form of controlled association, it is a combining of past experiences in order to solve problems which cannot be solved by mere reproduction of earlier solutions. At the same time the role of *direction* in reasoning and recall is emphasized, but how this factor operates with trial and error, past associations, and mere reproduction of earlier solutions to problems is not faced. We are here smack up against the problem of organization which many workers in biology as well as psychology realize cannot be dealt with adequately except as a problem in its own right. From the reviewer's experience such problems cannot be evaded even in the beginning course because many students have already faced them in courses in philosophy, logic, biology, and elsewhere.

The section on motivation of behavior seems to this reviewer to be best in the opening chapter dealing with physiological drives such as hunger, thirst, and sex where the material is largely drawn from experimental sources. The chapter on common social motives reads too much like a re-wording of the instinct psychology with too little use made of laboratory findings relevant to the topic. The chapter on conflict opens with sources of conflict in the environment and in the individual and then presents topological representation of conflict situations as an "interesting and illuminating method of representing and analyzing conflict situations" (p. 245). However, in the succeeding treatment of reactions to conflict such as compensation, identification, phantasy, projection, repression, and experimentally produced conflicts there is no further use made of Lewinian concepts. Again the integration must either be made by the instructor or the student will suffer from intellectual indigestion. Other alternatives are to omit topological representation or to put it at the very end of the chapter, pointing out that some, much, or most of the material discussed (depending upon the degree to which the instructor knows topological psychology) can be diagrammed in these terms. The author's penchant for trial and error pops out again in his recommendation of it as a possible solution in the alleviation or cure of conflict, contrary to the usual emphasis on rational procedures in psychotherapy. Since the patient admittedly knows why he is trying various lines of action, namely to find a way out of his conflict, it is doubtful if the procedure recommended is truly trial and error as Munn says.

The section on feeling and emotion which follows the one on motivation of behavior might well have preceded it, as affective states have been regarded by almost everyone as motivators of behavior. The main findings in the field are well covered with one or two exceptions. In the discussion of the Cannon-Bard theory the inhibitory function of the cortex is not mentioned and in the diagram illustrating the contrasting features of this theory as against the James-Lange theory the cortico-thalamic inhibitory path is not even shown. In view of the great im-

portance of the role played by the cortex in inhibiting emotion through positive inhibitory regulation and in allowing emotional expression through release of inhibition, the account offered here is entirely inadequate if not misleading, as reference to Bard's exposition in the *Handbook of General Experimental Psychology*, pp. 305–307, will show. Both text and diagram need this aspect of the theory for a correct as well as complete statement.

The following section, *Knowing Our World*, deals with attention and the special senses. Munn has here done an excellent job of boiling down the classical, and for the most part, stereotyped material and he has made it attractive by the use of well-chosen diagrams and half-tones. In view of the tremendous use made during the war years and now of material from the fields of sensation, perception, and psychophysics, not to mention their interrelations with sensori-motor learning, the time is past when we can rest content with traditional accounts of these fields. There is a wealth of material not yet in any text which modifies the whole approach to sensory processes and bears on every other field of psychology which should form part of the elementary student's equipment. Why recent work in some fields finds its way almost at once into textbooks and equally important work in other fields must wait a generation or more is hard to understand. For example, the explanation given of constancy is naive in the extreme in the light of not-so-recent work. And the epoch-making contributions by Katz find no reflection in the treatment of vision even though they had been in print for 35 years at the time this book appeared.

Several inaccuracies in terminology and fact should be corrected in future editions, such as: brightness and lightness should not be used interchangeably, and unless film and aperture modes are distinguished it is impossible to appreciate their difference; the assumption that Hering "neutral gray" is a constant or a general phenomenon rather than a special case is not tenable in view of work by Koffka and others; the discussion of retinal mixture versus overlapping of lights is so unclear it is impossible to determine what is meant and if it is correct; the usual explanation of *Flor* contrast as due to softening or obliteration of contours is palpably wrong and needs to be supplanted by the correct explanation given by von Bezold in the early part of the present century; the interchangeable use of note and tone, so common in discussions of hearing, should be replaced by more precise terminology in which tone relates to hearing-experience and note to the printed symbol. Only one figure of sensory qualities, the double cone for vision, is given although the smell prism and the taste tetrahedron are just as good in their respective modalities. The value of 1/3 as the Weber fraction for temperature is altogether too large to be representative following Culler's work.

In general the chapters just considered rely too much, on the theoretical side, on past experience and similar explanations and suffer

from a lack of unifying principles by which intra- and inter-sensory material may be related as well as unified with other psychological processes. If, as admitted, principles of organization are effective in attending, are they not perhaps also of importance in perception, and taking a further step, in learning and thinking as well? Recognition of such principles might unify and simplify psychology for the beginner.

The seventh and final section of the book, *Individual Differences*, contains chapters on statistics, intelligence, aptitudes, and personality. The chapter on statistics, kept until this section as an "Introduction to Statistical Analysis of Individual Differences," might well come earlier, especially for use in courses with laboratory. However the chapter can be introduced as it stands in almost any part of the course so its actual position matters little. The other three chapters form a fitting close to the book, entirely in the spirit of the more experimental portions in being packed with concrete material. Intelligence is approached from the historical angle and the important question of heredity and environment is quite fully discussed. The discussion of factor analysis—including the fundamental factors found by Thurstone, and the illustratory material from test batteries—make this chapter unique for an elementary presentation and one of the finest things in the book.

Similarly the chapters on aptitudes and personality are extremely well presented and again demonstrate the author's ability to condense a large amount of fact into a relatively small compass. In the chapter on personality the discussion includes methods of approach such as case history, rating, paper and pencil tests, behavior tests, interviews, free association, dream analysis, and projective methods, and also physique and temperament, role of the endocrines, and abnormal states. The open, empirical treatment here is more acceptable because the subject is more familiar to the average student and personality as a concept already provides some structuration by which its data can be ordered.

Taking the book as a whole, what are its pros and cons? On the plus side it is an excellent text in so far as it provides a wealth of concrete factual data chosen from widely different sources both within and outside psychology proper. With some exceptions it represents present-day scientific psychology very fairly. The student should come aways from this text with respect for the scientific approach not because he has been told that it works best in various fields but because he has found that material obtained by scientific methods can be applied to many different life situations and leads to further fruitful discovery. If the author is unable to accept a theory *in toto* his criticism is so mild and fair that the student's respect for the theory as well as for psychology in general is in no wise diminished. This and the catholicity of Munn's approach should exert a very good effect on coming generations of psychologists. Too often personal or institutional loyalties lead even graduate students to belittle men and work done outside their own bailiwicks with effects

detrimental both to themselves and to psychology. This book should serve as an excellent corrective to this tendency.

On the negative side of an otherwise fine piece of writing and presentation must be noted the lack of an integrating and unifying point of view which has been pointed out in our previous discussion. This lack results in a looser and more disjointed treatment than is necessary in the light of present advances along various fronts. This is not meant to imply that Munn himself does not have a point of view. As we have seen, careful reading reveals that for him trial and error is the great principle operating in human behavior and a number of indications are present that he believes in what has been dubbed an "atomistic logic," *i.e.*, proceeding from "simples" to "complexes." But having brought into his discussion of conditioning the principle of relevance, into thinking the principle of direction, into some sensory experiences primitive organization, and having recognized other whole-properties as well, he is under obligation to apply them more generally where they are applicable or at least to square them with the fundamental principles he believes are operative. Perhaps he has done this and this reviewer has missed it. If so, then it is probable that most students will fail to see how it all goes together.

The tendency to over-simplify has already been pointed out with reference to certain neural diagrams but it occurs much more frequently in the textual discussion where it leads the author, in making points which are quite valid in themselves, to say things he cannot possibly mean as they stand. For example, in writing about the influence of animal experiments on the theoretical basis of psychology, there appears the remarkable statement: "After all, learning is learning and vision is vision whether it occurs in man or animal" (p. 10). But the vision of the most widely used laboratory animal, the white rat, is very different from that of man, from retina to higher cortical centers, and Munn later points out that "Insight is rare in animals, not quite so rare in children, and quite common in human adults" (p. 109), meaning to distinguish among kinds of learning. One finds too many statements like this which take a good deal of explaining to mitigate.

There can be little doubt the present book will set a pattern for future introductory texts. The double columns while providing a shorter reading line and more words per page also make possible wider spaces for illustrative material and marginal notes. The wealth of charts, diagrams, and pictures lessens the instructor's blackboard work and should prove a boon to places where laboratory work cannot be given. In this text psychology appears as a positive, if not positivistic, science. If it were possible to combine what Munn has done with more emphasis on methods and unifying principles we should be much nearer the perfect presentation of present-day psychology everyone desires.

HARRY HELSON.

Bryn Mawr College.

BRIDGES, J. W. *Psychology normal and abnormal.* Toronto: Sir Isaac Pitman & Sons, 1946. Pp. xviii+470.

Except for dropping the chapter on philosophical foundations, the splitting and partial revision of the chapter on reflexes and instincts, and the annotation of the extensive bibliography, the 1946 edition has "identical twin" resemblance to its 1930 predecessor. The appearance of the revision does, however, call attention again to a book with a classical timelessness of integration (despite an eclectic tolerance), stimulating hypotheses, and a style reminiscent of William James. The general reader and even the professional psychologist will find interest and value here, although the book was designed for the introductory psychology course of pre-medical and medical students.

Bridges chooses to give the distilled essence of a topic rather than to lead the reader to a conclusion from the raw data of experiments and case studies. The few graphs, tables, and other reference to specific studies are illustrative only. Just as the dramatist's words furnished the bare Elizabethan stage, the emphasis on the logic of the argument in this book seems to stimulate more associations and imagery than one gets from many texts replete with illustrations. Many more of the quotations are from English and French psychologists than one meets in most American texts.

The chapter headings might have come from any of a number of general psychologies, but the plan of devoting the first half of each chapter to normal behavior and the remainder to the related abnormalities makes a distinctive pattern throughout the book. Technical words are italicized and well defined. The chapter on applied psychology delimits that field with such precision and perspective that it ought to be widely reprinted.

An error not corrected from the 1930 edition is the taking of the standard deviation from the median. Emphasis on the older studies, e.g., Downey will-temperament tests, is heavier than on those of the last two decades.

The problem of how to teach psychology in medical schools or to pre-medical students seems to have led in at least three main directions. Some have emphasized a sociological-psychological approach, as in Pressey's *Life*, since this is a common medical blind spot. Others have stressed the genetic-psychosomatic attitude, as found in such authors as Maslow and Mittlemann. A third group believe that medical students are more highly motivated and gain more insight from the contrasts and comparisons of the normal and abnormal. Bridges from his experience as the first professor of abnormal psychology on a medical faculty has provided an effective text for the last group.

GEORGE M. HASLERUD.

University of New Hampshire.

GRAY, J. S. *Psychology in human affairs.* New York: McGraw-Hill, 1946. Pp. viii+646.

While this book is, in many respects, a successor to Gray's previously edited text, *Psychology in Use* (American Book Co., 1941), in that it discusses the applications of psychology to the main fields of practical life, and represents the co-authorship of eleven other contributors, nevertheless it is not merely a revision. With two exceptions, the co-authors are new. They are less well known than those of the former book, but Gray has himself taken a more active part in the actual writing of the text. Several chapters appear for the first time, such as "Psychology in Speech Correction," "Psychology in Music, Art, and Literature," and "Psychology in Military Affairs." Others appear under new titles, and are written from a new viewpoint.

Perhaps the outstanding characteristic of the book is its emphasis on factual material. For example, Chapter II, on "Psychology in College Life" contains twenty tables and four graphs. Chapter III, on "Child Development" contains ten graphs and twenty-one tables. Much of this material is new to textbooks, and with few exceptions, the references are to studies published after 1930. The general effect of this emphasis on experimental data and practical findings is to require a change in teaching methods on the instructor's part. His function is no longer to supplement the text with up-to-date illustrative material, but rather to interpret and evaluate that which is given. Less supplementary assigned reading is needed, and much more digesting of the text by the student. The art of reading and interpreting tables and graphs is one which requires special training. Many students are allergic to statistics, though this is not in itself an argument for using them sparingly, if the instructor is competent to vitalize them. But the chief value of facts is to illustrate and support laws, principles and theoretical formulations. They are most effectively used in the inductive development of a topic. Most of these facts are and should be promptly forgotten by the student, so that his memory is freed for the permanent retention of the principles. The immature student needs much expert guidance in recognizing the bare essentials of fact to be learned. While this book is many strides ahead of the type which presents only unsupported assertions, or illustrations selected for their patness only, does it err slightly in the opposite direction?

Another important characteristic of the book is its emphasis on the practical. This is to be expected in an applied psychology text, but is seldom achieved. Omnibus books too often give an impression of sketchy remoteness, with little practical contact, while technical treatises are written for the advanced student who wants specialized information. This book, by achieving a compromise between these two extremes and by emphasizing the practical aspects of each field for the layman, fills a

real need. Its range of topics is wider, its treatment more complete than is customary in such books.

In spite of its up-to-dateness, the book occasionally presents old data which have been superseded, or theory which is now modified. In one or two cases, quite erroneous statements appear, such as the following, in connection with a discussion of the topic of I.Q. constancy, on pages 91–92:

> If the child develops mentally at exactly the same rate as other children tested, his I.Q. will remain constant. However, if his mental development is faster than that of other children, his I.Q. will increase. Likewise, if his mental development is slower than that of other children, his I.Q. will decrease.

The author seems to have confused *constancy* of I.Q. with *normality* of I.Q., for if the statement were taken as it stands, it would mean that no child's I.Q. is constant if he has a faster or slower developmental rate than the average child.

An innovation in this book which probably has pedagogical value and would be used oftener by authors of texts if publishers would let them, is the table of contents at the beginning of each chapter. The addition of page numbers would increase its value. A word must be said in criticism of certain reproduced charts, in which the reduction in size of print needed to get them on the page has made them unreadable; for example, those on pp. 474 and 573. Otherwise the style of the book is good.

Many psychology teachers will welcome this book either as a supplement to the general course in psychology, or as a second course to follow the introductory one. Those who teach adult extension classes will find it an excellent survey text, both meaty and comprehensive.

ARTHUR G. BILLS.

University of Cincinnati.

LUCK, J. M., & HALL, V. E. (Eds.). *Annual review of physiology* (Vols. VII & VIII). Stanford Univ. P.O.: Annual Reviews, Inc. and American Physiological Society, 1945, 1946. Pp. vi+774 (Vol. VII). Pp. vi+658 (Vol. VIII).

These are Volumes VII and VIII of the annual series begun in 1937 and published jointly by the American Physiological Society and Annual Reviews. It is the declared editorial policy of the *Review* that "encouragement is given only to preparation of reviews which survey the important contributions of the preceding year or biennium, which appraise them critically and evaluate with discrimination the present status of the subject. Comprehensive reviews in which the task of the author is one of compilation rather than of appraisal are deliberately eschewed." Despite this policy, some of the reviews are principally compilations or annotated bibliographies. And some are rather spotty

compilations mixed with evaluation, while only a few reach the goal of really critical reviews of the literature. On the whole, however, the reader can obtain a picture of the more significant aspects of research progress in the respective fields covered by the review.

Volume VII contains 26 chapters written by 30 authors, and Volume VIII contains 25 chapters by a total of 29 authors. In each case, bibliographies of literature cited in the various reviews total about 4,000 references. At the end of each volume is an author index of about 4,000 items and a subject index of about 40 pages in length.

Because the reviews are written by physiologists for physiologists, more than half of each volume is of no interest to the psychologist except that occasionally there is a brief treatment in a sentence or two of some psycho-physiological problem. The psychologist, however, who wants to find out what has happened recently in some special aspect of physiology relevant to his field of teaching or research will find that his best bet is to go to these volumes and to consult the excellent author and subject indices before resorting to other less up-to-date textbooks or to more laborious methods of library research.

More than that, however, many of the special chapters in physiology are good reading for psychologists engaged in the respectively related field of psychology: for genetic psychology, *Physiological aspects of genetics* VII and VIII) and *Developmental physiology* (VII and VIII); for sensory psychology, *The special senses* (VII) and *Audition* (VIII); for neural mechanisms of behavior, *Electrical activity of the brain* (VII), *Conduction and synaptic transmission in the nervous system* (VII), *Nerve and synaptic conduction* (VIII), *The visceral functions of the nervous system* (VII and VIII); and for a general review of physiological psychology, *Physiological psychology* (VII and VIII).

The contrast between the chapters on physiological psychology in the two volumes deserves a special comment. In Volume VII, Stone has presented a careful and critical review. He has covered thoroughly the recent literature and has appraised its strength and shortcomings so that the reader can see what has happened and what it means for physiological psychology.

The same chapter in Volume VIII by Seashore, however, does neither of these two things. It opens with a philosophical discussion of the mind-body problem and of the scientific approach to it. The chapter then proceeds to summarize the present status of individual differences in skills, abilities, aptitudes and capacities. Finally, it gives a general summary of the effects of extreme working conditions upon the effectiveness of human performance. Thus Seashore's chapter spends a lot of time on problems which are not physiological psychology, in any reasonable definition of the field; and he fails to review or appraise the recent literature in the field. As a consequence, the physiologist reading the two chapters is likely to be bewildered by two so very different concepts of physiological psychology.

Looked at in perspective, these two volumes of *Annual Review of Physiology*, like previous volumes in the series, are an extremely valuable aid, not only to physiologists, but to all those for whom physiology is an important ancillary subject. By and large, the chapters give scholarly up-to-date appraisals of their respective fields. As he has stated before, this reviewer feels that a companion volume, giving an annual review of psychology, would be an invaluable aid to psychologists, which would help us "keep up with the literature" and give us better perspective on the developments in our field.

CLIFFORD T. MORGAN.

The Johns Hopkins University.

BARKER, ROGER G., WRIGHT, BEATRICE A., AND GONICK, MOLLIE R. *Adjustment to physical handicap and illness: a survey of the social psychology of physique and disability.* New York: Social Science Research Council, 1946. (Bulletin No. 55.) Pp. xi+372.

This is another in the excellent series of research summaries sponsored by the Social Science Research Council. The authors have earned special commendation by providing intelligently critical comments as to the assumptions and thinking of earlier investigators, rather than merely reporting data and conclusions; by writing this summary of prior research into a theoretical frame of reference (topological psychology); and by introducing some well-chosen original material to illuminate the inferences drawn from published sources.

From their survey the authors have eliminated somato-psychological studies of age, sex, race, and speech defects, on the ground that these have recently been covered adequately by other reviewers. Leprosy is discarded as a minor problem in the western world. Of the remaining areas, detailed reports are presented on: normal variations in physique; crippling; tubercular conditions; impaired hearing; and acute illness. Bibliographies are added on: visual disability; cardiac conditions; diabetes mellitus; cosmetic defect; rheumatism; and cancer.

The least satisfactory chapter is that on variations in normal physique. There is a good section on size changes at adolescence, but the discussion of variations in adult size is rather elementary, and no mention whatever is made of Sheldon's *The varieties of human physique* and *The varieties of temperament.* Even if one does not accept Sheldon's theory, his work can hardly be ignored. The authors occasionally dip a hesitant toe into the cold waters of endocrinology, genetics and autonomic nervous function, then withdraw hastily. Clarity would have been improved by frankly excluding such material.

Outstanding treatments of crippling, tubercular conditions and impaired hearing more than atone for any shortcomings of the earlier chapter. Particularly interesting is the mode of analysis in terms of *overlapping situations.* A disabled person is able to function on a par with normals in some environments; he is decisively barred from other

situations; but between these extremes will fall a range of ambiguous conditions in which the handicapped person may participate, but under difficulties. The Lewinian concepts of barrier, valence, potency and congruence are used fruitfully to show basic similarities between the situations facing the orthopedic cripple, the tubercular, the deaf and the individual with acute illness.

The person with impaired hearing, for example, functions in many situations unnoticed by his normal associates. If the behavior involved does not require auditory controls, he may compete on equal terms. Where hearing is involved, he may be handicapped and subject to extra criticism, since his impairment is not obvious and many normals (e.g., school teachers) fail to make allowances for it. The barriers in his field are indefinite (as contrasted with the orthopedic cripple, for example, to whom certain activities are plainly impossible), and this condition often gives rise to vacillation and instability. The valence of full-normal activity is positive and high, but the valence of failure and criticism is negative and high. Thus physically handicapped persons are likely to show the familiar symptoms of conflict.

We occasionally feel, in these topological analyses, that there is an unstated shift from the topology of the external situation (geographical environment) to the situation as perceived by the individual (behavioral environment). In the case of cripples, for example, some activities are objectively impossible, whereas others are subjectively considered to be impossible. It is clear that these two should not be treated as identical, and yet that impression is sometimes given. If the entire analysis were erected on a perceptual basis, this uncertainty could have been avoided.

The importance of the individual's perception of his defect, and of his behavioral field, is well illustrated by the discussion of family attitudes toward the handicapped. Many parents reject the handicapped child, while others over-protect him and keep him in an infantile status. Optimum adjustment seems to be achieved when the parents adopt an understanding, objective attitude which focuses the child's attention on realistic assessment of the situation. Excessive sympathy and pity are likely to encourage exaggeration of barriers and exclusion of many possibilities for normal participation.

The final chapter on employment of disabled persons gives a realistic and well-considered treatment of this problem, the solution of which is basic to optimum adjustment of handicapped adults.

ROSS STAGNER.

Dartmouth College.

LEWIS, CLAUDIA. *Children of the Cumberland.* New York: Columbia Univ. Press, 1946. Pp. xviii+217.

Before going to the Southern highlands, Miss Lewis was a teacher in the Harriet Johnson Nursery School in Greenwich Village,

City. She compares the behavior of the children in the nursery school which she established in the mountains of Tennessee with the behavior of her Greenwich Village pupils. The majority of the material presented the book concerns the mountaineer subjects.

A considerable part of the volume consists of a collection of incidents involving child care or child behavior. These range in length from single sentences to a page or more, in age of subjects from birth to senility, in form from dialogue to descriptive essays. They are extremely readable and serve to render very vivid the life of the mountain people.

Miss Lewis does not attempt to present quantitative measures, but for this she cannot be reprimanded. It is apparent that she devoted very full days to the nursery school, and that research had a secondary place. Nevertheless, her thinking is quantitative. She emphasizes the diversity of individual behavior which takes place in both schools, and makes clear that there is overlapping between the schools. However, it is her belief that there are large differences in central tendencies between her two kinds of subjects. With this contention, it is likely that nearly every person who is familiar with both types of children will agree.

Miss Lewis finds more spontaneity, more energy, more conflict, more aggression in the New York group. The mountain children are more placid, more compliant, more quiet, and in some respects, better adjusted.

She does not find the explanation of these differences in any one factor. Among the probable causes mentioned are the following: the differences in spaciousness of the environment, differences in climate, health, and nutrition, differences in sleeping habits, differences in infant care and family structure, differences in discipline, and differences in environmental stimulation.

Throughout the book, Miss Lewis shows an excellent understanding of child development in both New York and Tennessee—not an easy achievement. She also displays a high degree of ability as a writer. This combination of traits makes her book one of the best in the "child in a culture" field. It should be profitable alike to teacher, parent, psychologist, and sociologist.

The reader may sample for himself some of Miss Lewis's attitudes and style in the following quotation from her concluding chapter:

> No, now that this study is made I am not packing my trunks with the intent of moving down to Tennessee, building me a cabin, taking to the "simple life" and rearing my hypothetical children in the way the Summerville families do. For us it is not a question of attempting to turn the clock back in that way, which, indeed, would be as impossible as it would be undesirable. It is rather a question of trying to bring to Greenwich Village a little more of . . . Summerville life . . . .

WAYNE DENNIS.

University of Pittsburgh.

LEEPER, ROBERT. *Psychology of personality.* Ann Arbor: J. W. Edwards, 1946. Pp. 167.

The format of this book, resembling that of a laboratory manual or workbook, may lead many to overlook this significant treatment of personality and mental hygiene.

The author's organization apparently proceeds from an assumption which has increasingly impressed itself on the reviewer in recent years: namely, that, in order to have functional significance, any treatment of mental hygiene must be based on a consistent theory of personality processes. Moreover, such a treatment should not be left among the author's unverbalized potentialities, but should be given systematic formulation. It is to this task that the major emphasis of Leeper's book is devoted.

Leeper's treatment is mainly an elaboration of the following thesis:

> In general, the term "personality" covers three things: (1) the person's motives, and especially his emotional motives, or ways in which he responds emotionally in different life situations; (2) the general techniques by which, characteristically, he tries to attain satisfaction for these motives; and (3) the background of meanings or pictures of reality which determine the motives and types of adjustive responses of the person (p. 5).

The discussion of motivation distinguishes between physiological and emotional motives and between positive can negative emotional motives. While the latter distinction seems arbitrary, since the "negative" motives can be regarded as the products of frustration of the "positive" motives, it serves a useful expository purpose when the author deals with motivational differences existing between well adjusted and poorly adjusted personalities.

The techniques by which the person tries to attain satisfaction for his motives are considered from two points of view: (1) the nature of the learning processes, and (2) the description of effectual and ineffectual adjustment techniques. The learning processes are treated with due attention to the dynamic complexities recognized in modern learning theory. In addition to describing the usual techniques employed by maladjusted personalities, the author discusses some of the major techniques by which superior personalities distinguish themselves.

In his discussion of the "background of meanings," the author deals competently with an aspect of behavior which seldom receives the emphasis merited by its significance for personality dynamics. Leeper's thesis is that " . . . a person cannot govern his behavior just by what is objectively and actually true, but . . . must forever live and react in terms of properties which he infers as existing because of his experience in previous situations" (p. 92). This view that behavior is determined not as much by the character of objective reality as by the individual's interpretation of reality (through the phenomenon of emotional transference) is supported in terms of the principle of equivalence of stimuli and the principle of substitute response or displacement. Treated in

these terms, the "background of meanings" is seen to be an aspect of personality whose importance has been emphasized by such widely separated disciplines as the research of animal psychologists and the clinical observations of psychoanalysts.

It is the reviewer's opinion that through the medium of Leeper's book the principles of personality functioning are made understandable to the average undergraduate without undue simplification or loss in organic quality. For this reason, it seems regrettable that the book was not produced in a form more likely to have wide distribution.

The book will probably be disappointing to those who feel that a textbook should serve as a compendium of psychological research findings. While the author draws freely upon research sources, these tend to lose their distinctive identities in the author's discussion. No bibliography or index is provided.

BERT R. SAPPENFIELD.

Montana State University.

KELLEY, DOUGLAS M. *22 cells in Nuremberg.* New York: Greenberg, 1947. Pp. 245. $3.00.

The author of this book was for five months the official psychiatrist at the Nuremberg prison and in that capacity made psychiatric examinations af all the 22 top-ranking Nazi prisoners. The customary medical and psychiatric procedures were supplemented by Rorschach personality tests and Wechsler-Bellevue intelligence tests given by the author's fellow officer, Capt. G. M. Gilbert. The examinations and tests were further supplemented by information obtained from former intimate associates of the accused and from motion pictures, speeches, writings, and other records.

Except for Rudolf Hess and Hermann Goering, this was the first psychiatric study to be made of any of the accused. In view of the fact that twelve of the group are no longer living and that all but three of the others were disposed of by prison sentences of from ten years to life, the documents and conclusions of Dr. Kelley are destined to be of lasting historical interest.

The task which the author set himself was not merely or chiefly to determine the degree of mental responsibility of the subjects, but rather to investigate their basic personality patterns. He wanted to find out what these men were like who had made themselves masters of eighty million people, and what factors in childhood, youth, and later years had made them what they were. The book attempts to answer these questions in language sufficiently nontechnical to be intelligible to readers who are relatively unfamiliar with esoteric theories of personality. We are informed that more detailed reports of the work will be published later in professional journals, and that transcripts of the interviews and other records will ultimately be available to historians.

Among those to whom most space is devoted are Goering (27 pages), Hess (22 pages), Ley (21 pages), Rosenberg (13 pages), and Streicher

(11 pages). The other 17 members of the group get from three to eight pages each. By good luck the examination of Ley had been completed before he committed suicide, and we are told that the post-mortem examination of his brain confirmed the psychiatric diagnosis.

With the exception of one chapter, the book is concerned entirely with the 22 Nazis who were studied first-hand by the author. The additional chapter presents a 35-page portrait of Hitler based on information and comments obtained from the Führer's contemporaries, his aides, his personal physicians, and his secretaries. Some of this information is new, and the author's interpretations differ in several important respects from those which have been current.

Within the limits of a brief review it is not possible to summarize the author's interpretations of the individual subjects. In fact, each portrait as sketched is a unified gestalt that almost defies further condensation. The sitters for these portraits composed indeed a motley group. They ranged from the stupid to the highly intelligent; from the semi-insane to the stable and well integrated; from the shrewd and talented leader to the errand-boy hanger-on seeking in Hitler a father surrogate. But there were three characteristics which they had in common: inordinate ambition, debased ethical standards, and a hyperdeveloped nationalism that justified anything done in the name of Germandom—plus, of course, an economic and political environment that allowed full play to their ruthless wills.

The author's conclusion is that Nazism was a "socio-cultural dis-ease," epidemic among our enemies but endemic everywhere. He tells us that the Nazi leaders were not the rare and spectacular types that can be expected to appear only once in a century. Instead, neurotics like Hitler, with "hysterical disorders and obsessive complaints, can be found in any psychiatric clinic." Similar ones, thwarted and discouraged, but determined to do great deeds, roam the streets of every American city. "Strong, dominant, aggressive, and egocentric personalities like Goering . . . can be found anywhere—behind big desks deciding big affairs as businessmen, politicians, and racketeers." We hardly need to be reminded that men strongly resembling some of these types occasionally win election to our highest law-making bodies or to the governorship of a great state.

Dr. Kelley has analyzed for us 22 types of totalitarian-virus, has described the soil in which they thrive, and has indicated some of the means by which society can protect itself against them. His book will inevitably be compared with one written by another psychiatrist—Brickner's *Is Germany Incurable?* Of the two, the reviewer finds Kelley's less controversial and no less challenging.*

LEWIS M. TERMAN.

Stanford University.

* Since this review was written, *Nuremberg Diary*, by Capt. G. M. Gilbert, has been published. This book should be read along with Dr. Kelley's. L.M.T.

BOOKS AND MATERIALS RECEIVED

ALLPORT, G. W., AND POSTMAN, LEO. *The psychology of rumor.* New York: Henry Holt, 1947. Pp. vii+247.

ALSCHULER, ROSE H., AND HATTWICK, LA BERTA W. *Painting and personality.* (2 Vols.) Chicago: Univ. of Chicago Press, 1947. Pp. xi+590.

AXLINE, VIRGINIA MAE. *Play therapy.* Boston: Houghton Mifflin, 1947. Pp. xii+379. $3.50.

BARUCH, DOROTHY, AND MONTGOMERY, ELIZABETH. *The girl next door.* Chicago: Scott, Foresman, 1947. Pp. 256.

BAUMGARTEN-TRAMER, F. *Der Rorschach—Test im Lichte der Experimentellen Psychologie.* Archivio di Psicologia Neurologia e Psichiatria, 1946, **8**, No. 2. Pp. 37.

BECK, H. P. *Men who control our universities.* New York: King's Crown Press, 1947. Pp. x+229.

BOWERMAN, WALTER G. *Studies in genius.* New York: Philosophical Library, 1947, Pp. 343.

CARTER, L. F. (ED.) *Psychological research on navigator training.* Army Air Forces Aviation Psychology Program, Research Reports. Report No. 10. Washington: U. S. Government Printing Office, 1947. Pp. iii+186.

COLE, LUELLA, AND MORGAN, J. J. B. *Psychology of childhood and adolescence.* New York: Rinehart, 1947. Pp. xi+416.

COUNT, E. W. *Brain and body weight in man: their antecedents in growth and evolution.* Annals of the New York Academy of Sciences, 1947, Vol. XLVI, Art. 10. Pp. 993–1122.

COWLES, E. S. *Don't be afraid.* Chicago: Wilcox & Follett, 1947. Pp. xv+254.

DAVIS, F. B. (ED.) *The A. A. F. qualifying examination.* Army Air Forces Aviation Psychology Program, Research Reports. Report No. 6. Washington: U. S. Government Printing Office, 1947. Pp. iii+266.

DAVIS, F. B. *Utilizing human talent.* Washington: American Council on Education, 1947. Pp. ix+85.

DICHTER, E. *The psychology of everyday living.* New York: Barnes & Noble, 1947. Pp. x+239.

ERICKSON, C. E. (ED.) *A basic text for guidance workers.* New York: Prentice-Hall, 1947. Pp. x+566.

FALVEY, HAL. *Ten seconds that will change your life.* Chicago: Wilcox & Follett, 1947. Pp. 96.

FIELD, HARRY. *Midiendo la Opinion Publica.* Mexico, D. F.: In-

stituto de Estudios de Psicologia Social y Opinion Publica, 1945. Pp. 56.

FORD, MARY. *The application of the Rorschach test to young children.* Institute of Child Welfare Monograph, No. 23. Minneapolis: Univ. of Minnesota Press, 1946. Pp. xii+114.

FREDERICK, R. W., KITCHEN, P. C., AND MCELWEE, AGNES R. *A guide to college study.* New York:Appleton-Century, 1947. Pp. viii+341.

GARRETT, H. E. *Statistics in psychology and education.* (3rd Ed.) New York: Longmans, Green, 1947. Pp. xii+465.

GILBERT, G. M. *Nuremberg diary.* New York: Farrar, Straus, 1947. Pp. 471.

HAGAN, W. A., *et al. The relation of diseases in the lower animals to human welfare.* Annals of the New York Academy of Sciences, 1947, Vol. XLVIII, Art. 6. Pp. 351-576.

HALL, R. B. *Area studies: with special reference to their implications for research in the social sciences.* New York: Social Science Research Council, 1947. Pp. iii+90.

HALL, V. E. (Ed.) *Annual review of physiology* (Vol. IX, 1947). Stanford Univ. P.O.: Annual Reviews, Inc. & American Physiological Society, 1947. Pp. vii+736.

JELLINEK, E. M. *Recent trends in alcoholism and in alcohol consumption.* New Haven: Hillhouse Press, 1947. Pp. 42.

KITAY, P. M. *Radicalism and conservatism toward conventional religion. A psychological study based on a group of Jewish college students.* Teachers College, Columbia Univ., Contr. to Educ., No. 919. New York: Bureau of Publications, Teachers College, Columbia Univ. 1947. Pp. vii+117.

KRACAUER, S. *From Caligari to Hitler.* Princeton: Princeton Univ. Press, 1947. Pp. xi+361.

LE CRON, L. M., AND BORDEAUX, JEAN, *Hypnotism today.* New York: Grune & Stratton, 1947. Pp. v+278.

LEIGHTON, DOROTHEA, AND KLUCKHOHN, CLYDE. *Children of the people.* Cambridge: Harvard Univ. Press, 1947. Pp. viii+277.

LEPLEY, W. M. *Psychological research in the theaters of war.* Army Air Forces Aviation Psychology Program, Research Reports. Report No. 17. Washington: U. S. Government Printing Office, 1947. Pp. iii+201.

LINDNER, R. M., AND SELIGER, R. V. *Handbook of correctional psychology.* New York: Philosophical Library, 1947. Pp. 691.

LINK, H. C. *The rediscovery of morals.* New York: E. P. Dutton, 1947. Pp. 223.

LONDON, L. S. *Libido and delusion.* (2nd Ed.) Washington: Mental Therapy Publ., 1946. Pp. xi+259. $3.50.

LOUTTIT, C. M. *Clinical psychology.* (Rev. Ed.) New York: Harper, 1947. Pp. xviii+661.

LUNDBERG, G. A. *Can science save us?* New York: Longmans, Green, 1947. Pp. 122.

LUNEBERG, R. K. *Mathematical analysis of binocular vision.* Princeton: Princeton Univ. Press, 1947. Pp. vi+104. $2.50.

MICHOTTE, A. *La perception de la causalité. Études de Psychologie,* 1946. Vol. VI. Pp. viii+296.

NACHMANSOHN, D., *et al. The physico-chemical mechanism of nerve activity.* Annals of the New York Academy of Sciences, Vol. XLVII, Art. 4. Pp. 375–600.

NICOLE, J. E. *Psychopathology. A survey of modern approaches.* (4th Ed.) Baltimore: Williams & Wilkins, 1946. Pp. vii+268. $4.75.

RADKE, MARIAN J. *The relation of parental authority to children's behavior and attitudes.* Institute of Child Welfare Monograph, No. 22. Minneapolis: Univ. of Minnesota Press, 1946. Pp. x+123.

ROBINSON, F. P. *Effective study.* New York: Harper, 1946. Pp. ix+262.

SADLER, W. *Mental mischief and emotional conflicts.* St. Louis: C. V. Mosby, 1947. Pp. 396.

SANDOW, A., *et al. Muscular contraction.* Annals of the New York Academy of Sciences, 1947, Vol. XLVII, Art. 6. Pp. 665-930.

SHERIF, MUZAFER, AND CANTRIL, HADLEY. *The psychology of ego-involvements.* New York: John Wiley, 1947. Pp. v+525.

SNYDER, W. U. (ED.) *Casebook of non-directive counseling.* Boston: Houghton Mifflin, 1947. Pp. viii+339. $3.00.

SOROKIN, P. A. *Society ,culture and personality: their structure and dynamics.* New York: Harper, 1947. Pp. v+742.

THORNDIKE, R. L. (ED.) *Research problems and techniques.* Army Air Forces Aviation Psychology Program, Research Reports. Report No. 3. Washington: U. S. Government Printing Office, 1947. Pp. viii+163.

THURSTONE, L. L. *Multiple-factor analysis.* Chicago: Univ. of Chicago Press, 1947. Pp. xix+535.

WOLFF, WERNER. *What is psychology?* New York: Grune & Stratton, 1947. Pp. vii+410.

WOODWORTH, R. S., AND MARQUIS, D. G. *Psychology.* (5th Ed.) New York: Henry Holt, 1947. Pp. iii+677.

WORTIS, S. B., *et al. Physiological and psychological factors in sex behavior.* Annals of the New York Academy of Sciences, 1947, Vol. XLVII, Art. 5. Pp. 603-664.

L'Anneé Psychologique. Henri Piéron (Ed.). Forty-third and forty-fourth years (1942-1943). Paris: Presses Universitaires de France, 1947. Pp. 856.

Archivio di Scienza della Cerebrazione e dei Psichismi. M. Levi—Bianchini (Ed.). Vol. I–II, 1944–45. Pp. 105–208.

1946 fall testing program in independent schools and supplementary studies. Educational Records Bulletin, No. 47. New York: Educational Records Bureau, 1947. Pp. x+58.

Revista de Psicologia General y Aplicada. Vol. 1, No. 1. Madrid: Instituto Nacional de Psicotecnia, 1946. Pp. 316.

Revista do Centro Psiquiatrico National. Vol. I, No. 1. Rio de Janeiro, Brazil: Imprensa Nacional, 1946. Pp. 161.

Theoria. Ake Petzall (Ed.). Vol. XII, Parts I–II. Lund, Sweden: C. W. K. Gleerup, 1946. Pp. 133.